UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Sciences de l'Information et de la Communication**

Arrêté ministériel : 7 août 2006

Présentée par

Marilyne LATOUR

Thèse dirigée par **Laurence Balicco**

préparée au sein du **Laboratoire GRESEC (EA 608)**
dans **l'École Doctorale Langues, Littératures et Sciences Humaines**

Du besoin d'informations à la formulation des requêtes :
étude des usages de différents types d'utilisateurs visant l'amélioration d'un système de recherche d'informations

Thèse soutenue publiquement le **24 juin 2014**
devant le jury composé de :

Mme Laurence BALICCO
Professeur, Université Stendhal Grenoble 3, Directrice
Mme Sylvie LAINÉ-CRUZEL
Professeur, Université Jean Moulin Lyon 3, Président
Mme Josiane MOTHE
Professeur, Université Paul Sabatier Toulouse 3, Rapporteur
Mme Céline PAGANELLI
MCF-HDR, Université Paul Valéry Montpellier 3, Membre

Université Joseph Fourier / Université Pierre Mendès France /
Université Stendhal / Université de Savoie / Grenoble INP

RÉSUMÉ

Devant des collections massives et hétérogènes de données, les systèmes de RI doivent désormais pouvoir appréhender des comportements d'utilisateurs aussi variés qu'imprévisibles. L'objectif de notre travail est d'évaluer la façon dont un même utilisateur verbalise un besoin informationnel à travers un énoncé de type « expression libre » (appelé langage naturel) et un énoncé de type mots-clés (appelé langage de requêtes). Pour cela, nous nous situons dans un contexte applicatif, à savoir des demandes de remboursement des utilisateurs d'un moteur de recherche dédié à des études économiques en français. Nous avons recueilli via ce moteur, les deux types d'énoncés sur 5 années consécutives totalisant un corpus de 1398 demandes en langage naturel et de 3427 requêtes. Nous avons alors comparé l'expression en tant que tel du besoin informationnel et mis en avant ce qu'apportait, en termes d'informations et de précisions, le recours à l'un ou l'autre du langage utilisé.

Mots-clés

Recherche informations ; Besoin informationnel, Expression et interprétation des besoins ; Formulation question ; Langage naturel ; Termes recherche ; comportement utilisateur

Abstract

With the massive and heterogeneous web document collections, IR system must analyze the behaviors of users which are unpredictable and varied. The approach described in this paper provides a comparison of the verbalizations for both natural language and web query for the same information need by the same user. For this, we used data collected (*i.e.* users' complaints in natural language and web queries) through a search engine dedicated to economic reports in French over 5 consecutive years totaling a corpus of 1398 natural language requests and 3427 web queries. Then, we compared the expression of the information need and highlighted the contributions in terms of information and clarification, the use of either language used.

English Keywords

Information retrieval ; Information Need, Query formulation and query expression ; Query formulation ; Natural language ; Search term ; User behavior

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

CLEF	Cross-Language Evaluation Forum
CTS	Covering Topic Score
EN	Entity Recognition
IHM	Interface Homme-Machine
LN	Langue Naturelle
LSA	Latent Semantic Analysis
MEDLARS	MEDical Literature Analysis and Retrieval System
NER	Named-entity recognition
NLP	Neuro-linguistic programming
NTCIR	National Test Collection for IR Systems
RI	Recherche d'Informations
SGML	Standard Generalized Markup Language
SMART	System for the Mechanical Analysis and Retrieval of Text
SRI	Système de Recherche d'Information
STAIRS	STorage And Information Retrieval System
TAL	Traitement Automatique des Langues
TALN	Traitement Automatique des Langues Naturelles
TREC	Text REtrieval Conference
XeLDA	Xerox Linguistic Development Architecture

L’Homme raisonnable s’adapte au monde, tandis que l’homme déraisonnable persiste à vouloir adapter le monde à lui-même écrivait Bernard Shaw. C’est pourquoi tout progrès dépend de l’homme déraisonnable.

Pour ma *maman*
Elle aurait très fière.

Pour mon *papa* et mon *frère*
Ils m’ont soutenue et encouragée.

Pour *Tristan* et *Charlotte*,
Cette thèse a bercé leurs enfances.

Pour *Matthieu*,
mon amoureux.

REMERCIEMENTS

Au terme de ces années de thèse qui me conduisent aujourd'hui à présenter ce manuscrit, c'est avec un réel plaisir que je profite de cette section pour remercier celles et ceux qui ont été associés, de prêt ou de loin, à mes travaux de recherche.

Merci à ma directrice de thèse, Laurence Balicco, qui habituée au travail interdisciplinaire m'a apporté des pistes, des corrections, de la sérénité dans les moments de doute et surtout une grande liberté de recherche et de collaborations.

Je remercie chaleureusement mesdames Sylvie Lainé-Cruzel et Josaine Mothe d'avoir accepté d'évaluer cette thèse ainsi que Mme Céline Paganelli qui a bien voulu être examinateur, leur expertise en tant que références dans la communauté des sciences de l'information et de la communication, de la recherche d'informations est très précieuse.

Je remercie ceux auprès desquels j'ai contracté, faute de disponibilité, de temps ou de présence, des dettes morales immenses : ma famille et plus particulièrement mes deux enfants, qui sont nés en cours de thèse et mes amis qui m'ont toujours entouré et soutenu.

Enfin, je remercie mon amoureux de m'avoir porté durant la bien trop période de rédaction de ce manuscrit, et plus globalement de partager ma vie et mes rêves.

INTRODUCTION

Le 26 septembre 2013, *Google* annonce lors d'une conférence de presse pour fêter ses 15 ans, son nouvel algorithme baptisé *« Hummingbird »*. Ce nouvel algorithme s'éloigne de la logique de la recherche d'informations (RI)[1] pour s'ouvrir aux requêtes en langage naturel (LN). C'est un changement majeur pour le géant de la recherche dont l'objectif affiché est d'être capable de traiter des requêtes plus complexes et plus longues tout en prenant en compte le sens des mots dans leur contexte. Ce nouvel algorithme, en place depuis le mois d'août, est annoncé comme traitant actuellement 90 % des requêtes effectuées sur le site de *Google* à travers le monde ; les 10 % restant étant encore traité par l'ancien algorithme[2]. Il est cependant encore trop tôt pour évaluer des véritables répercussions d' *« Hummingbird »* sur les requêtes.

Si l'on reprend l'exemple de Danny Sullivan[3] à la requête « Quel est l'endroit le plus près de chez moi pour acheter un *iPhone* 5S ? », *Google* devrait pouvoir prendre en compte et lier les notions « endroit près de chez moi », « acheter », « iPhone 5S » sous-entendu : l'objet désiré, le type d'action exercée sur l'objet (celui d'acheter) et la couverture géographique (à proximité de là où est localisé l'internaute « chez moi »). L'objectif est donc double : *(a)* comprendre la demande dans sa globalité pour lui donner un « sens » plus exact et *(b)* capitaliser d'autres informations que celles apparaissant sur les pages de recherche. Plus largement et au travers de la demande exprimée, c'est l'interprétation du besoin informationnel de l'utilisateur qui est visé pour le premier point. Ainsi, dans l'exemple cité, cela reviendrait à comprendre que la demande concerne le domaine de la téléphonie mais aussi que le besoin s'étend également à une volonté d'acheter un appareil. Le second point, lui, s'intéresse au lieu d'habitation si l'internaute a déjà renseigné cette information. Également des requêtes précédemment effectuées sur le moteur peuvent aider à « contextualiser » la demande notamment sur les centres d'intérêts de l'utilisateur.

Ce nouvel algorithme rend compte d'une prise de conscience de la part des développeurs des systèmes de recherche d'informations (SRI) de traiter plus efficacement les requêtes avec une meilleure contextualisation du besoin informationnel.

---

[1] La RI recouvre les méthodes permettant de retrouver des informations dans un ensemble de documents au moment de leur indexation.

[2] Les Echos, « Google entame un big-bang pour son moteur de recherche », 27/09/2013.

[3] « FAQ : All About The New Google « Hummingbird » Algorithm », Danny Sullivan, 27/09/2013, disponible sur : http ://searchengineland.com/google-hummingbird-172816 .

Mieux répondre aux besoins informationnels

Le défi majeur qui se pose revient à la capacité de la machine à comprendre le besoin informationnel de l'utilisateur. C'est en effet un passage difficile aussi bien du côté de l'utilisateur que du côté système qui doit l'interpréter. Nous identifions plusieurs raisons à ces difficultés :

- Du côté utilisateur, une difficulté réside en premier lieu dans l'expression en tant que tel du besoin informationnel : l'utilisateur doit en effet avoir des connaissances sur ce qu'il lui manque afin de pouvoir le désigner. En second lieu, cet utilisateur doit pouvoir l'exprimer au travers d'un SRI et donc en maîtriser son utilisation.

- Du côté système, un premier obstacle réside dans l'appréhension du besoin informationnel à travers ce qui a été informatisé *via* le SRI. Cela induit donc de la part du système de faire des déductions sur le besoin informationnel à partir des comportements observables et des représentations extériorisées. La seconde difficulté côté système est de prendre en compte l'évolution du besoin informationnel. En effet, au cours d'une session de recherche, un besoin informationnel peut évoluer pour plusieurs raisons : *(1)* l'utilisateur a acquis suffisamment de connaissances pour formuler de façon plus adéquate sa demande *(2)* l'utilisateur a acquis plus de techniques documentaires pour mieux maîtriser les outils et les fonctionnalités de recherche d'un SRI *(3)* son besoin lui-même a pu se modifier en fonction des informations que l'utilisateur a obtenu ; un nouvel état de connaissances est né, modifiant ainsi le besoin informationnel initial.

Les moteurs actuels indexent des documents collectés sur le Web par des robots (*crawlers*) et proposent une interface utilisateur des requêtes permettant d'interroger la base d'index. Les documents sont au préalable « indexés » *i.e.* transformés par un ensemble de techniques permettant un accès facilité à l'information textuelle. L'appariement entre la requête et l'index va déterminer les documents à fournir en réponses à un besoin informationnel initial. De nombreux modèles de recherche d'informations (booléen, vectoriel, probabiliste) ont été testés *via* notamment des grandes campagnes d'évaluation comme TREC (*Text REtrieval Conference*), afin d'améliorer l'appariement entre les informations contenues dans des documents et les requêtes des utilisateurs. Or, à l'heure actuelle, cet appariement n'est toujours pas jugé satisfaisant pour plusieurs raisons : les utilisateurs emploient en moyenne 4 mots pour décrire leur besoin informa-

tionnel[4] ; créant ainsi certaines ambiguïtés sémantiques, les intentions des utilisateurs ne sont pas suffisamment prises en compte selon le contexte de la requête, l'environnement de l'utilisateur, le contexte des documents, le contexte des interactions. Dans cette logique, les modèles classiques de RI ne s'avèrent plus suffisants : basés sur une approche trop généraliste, ils répondent trop souvent des listes similaires de résultats pour des utilisateurs ayant émis une même requête mais ayant pourtant des besoins informationnels différents.

Compte tenu de ces limitations, la démarche de *Google* ainsi qu'une nouvelle génération de moteurs de recherche est d'exploiter le contexte de l'utilisateur ainsi que des connaissances liées à la requête. Cette démarche est celle de la Recherche d'Informations Contextualisée [ALL 02] dont le but est de mieux répondre aux besoins informationnels de l'utilisateur.

Mieux prendre en compte le « contexte »

Même si les questions liées au contexte du besoin informationnel ont été examinées depuis de nombreuses années, il y a pas de consensus sur la définition de ce concept [CAS 12] ; les points de vue ainsi et les facteurs qui permettent de construire le contexte sont divergents.

Rétrospectivement, les notions de *contexte* et *situation* ont tout d'abord été traitées de façon équivalente par [SAR 97] et [ING 96]. Le contexte (ou situation) y est défini(e) comme l'ensemble des facteurs cognitifs et sociaux ainsi que les buts et intentions de l'utilisateur au cours d'une session de recherche. [ALL 97] précise que le contexte couvre des aspects larges tels que l'environnement cognitif, social et professionnel dans lesquels s'inscrivent des situations liées à des facteurs tels que le lieu, le temps et l'application en cours.

Le contexte peut être défini comme « l'ensemble des circonstances particulières où un besoin d'information se pose » [MCC 99] : p. 58. Les éléments temporels et spatiaux sont alors particulièrement caractéristiques de la situation. [COO 01] soutient ce point de vue en définissant des situations comme étant des « environnements dynamiques, au

[4]Selon une étude menée par la société Chitika entre le 9 et le 12 janvier 2012, portant sur des centaines de millions de requêtes sur les moteurs de recherche *Google, Bing, Yahoo, Ask.com, AOL*. Le nombre de mots clés moyen tapés par les internautes sur *Google* serait de 4,29. Le moteur proposant les requêtes les plus « courtes » serait *AOL* (4,07) et les plus longues, *Ask.com* (4,81). Article disponible à l'adresse suivante :http ://www.abondance.com/actualites/20120126-11246-plus-de-4-mots-cles-en-moyenne-dans-les-requetes-moteurs.html

sein desquelles les processus d'interprétation se déroulent, se ratifient, changent et se consolident [COO 01] : p. 8.

[JUL04] met également en avant les constituants temporels et spatiaux en soulignant que « la situation dans la vie quotidienne commence par un événement ou un ensemble de circonstances qui créent pour une personne une prise de conscience d'un besoin d'informations » [JUL04] : p.547-548.

L'un des premiers exemples d'approches véritablement contextualistes aux besoins informationnels a été fournie par [WIL 81]. Il a souligné que les éléments contextuels qui affectent les besoins informationnels proviennent en particulier de rôles de travail des personnes dans les entreprises. Un rôle de travail peut être compris comme un ensemble d'activités et de responsabilités d'un individu. Les moyens par lesquels on peut satisfaire son besoin informationnel dépend de l'existence de barrières personnelles, sociales et de l'environnement (ibid : p. 10). Il place en cela l'utilisateur dans son contexte au centre du processus de recherche d'informations.

[GOK 02] préconise un modèle du contexte extensible et propose cinq éléments pour caractériser le contexte utilisateur : le contexte de l'environnement, le contexte personnel, le contexte de la tâche, le contexte social et le contexte spatio-temporel. Dans le même sens, [ING 05] définit six dimensions imbriquées du contexte : *(a)* la structure interne ou externe de l'objet (les unités documentaires qui le composent (des phrases d'un paragraphe ou des liens hypertextes) ; *(b)* les interactions et les activités qui se passent au cours d'une session ; *(c)* le contexte social ; *(d)* le contexte spatio-temporel ; *(e)* le contexte économique et technico-physique ; *(f)* le contexte de l'historique (actions passées effectuées par l'utilisateur lors d'une session de recherche).

Au-delà de ces multiples définitions, nous retiendrons plusieurs éléments qui caractérisent la notion de contexte : contexte de l'utilisateur, contexte de l'information, contexte de la tâche ou activité, contexte du système, contexte de l'environnement/physique, contexte temporel, contexte des interactions. Ces constituants sont importants en ce qu'ils rendent compréhensible la façon dont les besoins informationnels peuvent changer au sein et entre les situations. Mieux prendre en compte le contexte d'un besoin informationnel, c'est donc multiplier le recours à des disciplines variées : modélisation utilisateur, interaction homme-machine, cognition ou encore linguistique. Ce sont les travaux de [BEL 95] [ING 96] [SAR 97] qui amorcent notamment l'intégration des modèles cognitifs et des interactions dans la RI qui en étaient séparées.

Mieux croiser les approches disciplinaires

La recherche d'informations cherche des documents répondant à un besoin informationnel, exprimé à l'aide d'une requête.

Pour mieux contextualiser, personnaliser la demande et assister l'utilisateur dans sa démarche de RI, des travaux de divers horizons se sont focalisés sur la question du besoin informationnel :

- D'un point de vue cognitif : parce que celui-ci inclut un large éventail de processus mentaux mis en œuvre chaque fois qu'une information est reçue, stockée, transformée et utilisée. Les premières études sur le besoin informationnel traitent principalement de l'aspect cognitif : [WIL 96] avec la formalisation du comportement global de l'utilisateur, [KUH 91] avec le le modèle ISP (*Information Search Process*), Belkin [BEL 82a] avec l'état de connaissance « anormal » (*Anomalous State of Knowledge*). L'intérêt pour l'aspect cognitif du besoin informationnel a depuis diminué même si l'étude de l'aspect cognitif contribue à une meilleure compréhension de la phase initiale du processus de recherche d'information [SAV 12].

- D'un point de vue dialogue et interaction homme-machine (IHM) : les approches des besoins informationnels dans le cadre du dialogue et de l'interaction ont abouti à la conceptualisation qui soulignent le caractère spécifique et parfois unique de ces besoins. Les constituants de dialogue sont en effet variables , y compris des facteurs comme le sujet de conversation , le niveau de spécificité dans les questions d'articulation , de la terminologie utilisée et les rôles des participants à une conversation. Le besoin informationnel est également façonné par un processus de négociation qui peut faire l'objet d'une redéfinition. Le dialogue et les interactions illustrent un contexte dynamique dans lequel le besoin informationnel peut toujours rester ouvert à de nouvelles négociations.

- D'un côté linguistique : car celui-ci s'intéresse à l'étude du langage et au sens donné à un énoncé. Plus particulièrement, le Traitement Automatique de la Langue (TAL) ou Traitement Automatique de la Langue Naturelle (TALN) concernent l'ensemble des méthodes et des programmes qui permet un traitement informatisé des données langagières. Les méthodes classiques en RI effectuent déjà des traitements linguistiques comme la suppression des « mots vides » de sens (mots fréquents et/ou sans pouvoir discriminateur), la racinisation (réduction des mots de la même famille morphologique à une racine commune), la transformation du texte en « sac

de mots ». Ces simplifications effectuées à la fois sur les requêtes et les documents sont assez limitées ; car ne prenant pas ou peu le contexte pour l'interprétation des énoncés [JAC 00]. Des relations plus précises entre mots ou syntagmes sont utiles à l'interprétation des phrases. Notamment les relations grammaticales permettent de représenter la fonction des groupes de mots les uns par rapport aux autres. De même, des classes de mots (*i.e.* de même catégories sémantiques) regroupement des mots dont le sens est proche, ou des mots qui possèdent certaines propriétés sémantiques communes. Une meilleure prise en compte de la nature linguistique du matériau traité permet d'obtenir des améliorations des résultats de RI voire amène à d'autres applications que la simple recherche d'un document pertinent. C'est par exemple le cas avec les systèmes de Question-Réponses (Q/R) dont l'objectif est de fournir une réponse plus précise à une questions posée. Il existe une demande très forte sur le web pour des systèmes réalisant ce genre de tâche ; les dispositifs à mettre en œuvre doivent en tenir compte.

Contributions

Nous nous intéressons ici à la tâche de RI dont l'application la plus connue est celle de recherche par mots-clés dans les moteurs de recherche. Plus particulièrement, il nous est apparu que la prise en compte des informations du contexte de l'utilisateur pouvait constituer une piste d'amélioration intéressante pour l'ensemble des SRI. Notre intérêt principal s'est naturellement porté sur l'expression des besoins informationnels à la fois en langue naturel (correspondant ici à un langage d'« expression libre » et « sans contrainte ») et en langage de requêtes (sous forme de mots-clés *via* un formulaire de recherche d'un SRI). L'objectif de cette approche est d'évaluer la façon dont un même utilisateur verbalise un besoin informationnel à travers un énoncé de type langage naturel et un énoncé de type langage de requêtes.

En effet, la littérature actuelle est assez abondante en ce qui concerne les typologies de requêtes [BRO 02], [ROS 04], [JAN 08], l'analyse linguistique des requêtes [MOT 11], [PLU 11], les expansions de requêtes [TAN 05], [GAR 10], [VEN 09] notamment grâce à des ontologies de domaines [AIM 10] ou les dialogues homme-machine. A contrario, il n'existe pas, à notre connaissance, de travaux comparant les deux types d'énoncés précédemment cités. Ceci s'explique bien sûr par le fait que seulement dans de rares occasions l'utilisateur est appelé à formuler son besoin informationnel par ces deux types d'énoncés.

Notre étude expérimentale consiste à recueillir et à analyser les besoins informationnels exprimés à la fois en langage naturel et langage de requêtes par les mêmes utilisateurs, ceci dans un contexte bien particulier ; celui d'une demande de remboursement effectuée par des utilisateurs d'un moteur de recherche ayant exécuté des requêtes. Pour cela, nous nous situons dans un contexte applicatif, à savoir un moteur de recherche dédié à des études économiques en français. Nous avons recueilli *via* ce moteur de recherche, tous les besoins informationnels exprimés à la fois en langage naturel et en langage de requêtes sur 5 années consécutives (de 2002 à 2007). A travers un indicateur client (*logs*), nous avons pu recueillir à la fois son énoncé en langue naturelle (via un formulaire spécifique), la ou les requête(s) précédemment effectuée(s) dans le formulaire de recherche ainsi que des données personnelles (anonymisées). Nous avons totalisé un corpus de 1398 demandes en langue naturelle et de 3427 requêtes (une demande en langue naturelle pouvant être formulée par une ou plusieurs requêtes de la part de l'utilisateur).

Grâce à des règles linguistiques et une analyse morpho-syntaxique, nous avons schématisé l'énoncé en langage naturel, relevé des informations de contexte et obtenu une représentation sémantiquement comparable aux requêtes. Nous avons alors pu comparer les deux types d'énoncés et les distinguer en fonction des tâches de RI à accomplir.

Précisions que nos travaux concernent uniquement la manière dont sont nommés les besoins informationnels dans un premier temps de recherche d'informations. Ils ne prennent donc pas en compte les résultats obtenus ou les éventuelles reformulations.

L'objectif de cette thèse est donc double : *(i)* tout d'abord mieux connaître les comportements utilisateurs pendant une tâche de RI et *(ii)* les exploiter pour faciliter la compréhension des besoins informationnels *via* les SRI (notamment en fonction de la tâche à accomplir). Seront envisagés également les intérêts d'une telle démarche suivant le contexte de la recherche d'informations, le domaine et la tâche à effectuer de l'utilisateur ; ces aspects étant, comme nous l'avons déjà expliqué ci-dessus, généralement peu pris en compte voire ignorés par les SRI.

Organisation de la thèse

Nous effectuons dans la première partie un état de l'art de la littérature sur les besoins informationnels et leurs différentes modes d'expression au sein d'un SRI. Seront notamment analysés dans le **chapitre 1** les principes du besoin informationnel ainsi que les différents modèles cognitifs de RI. Dans le **chapitre 2**, nous étudions comment les SRI sont passés de paradigmes orientés « systèmes » vers des paradigmes orientés « uti-

lisateurs » et plus récemment des paradigmes plus spécifiquement orientés « tâches ». Nous examinons dans le **chapitre 3** les différentes formes d'expressions des besoins informationnels dans un SRI : langage naturel *versus* langage de requêtes. Nous mettons en lumière dans le **chapitre 4**, les études qui ont amorcé une analyse de la langue naturelle aux requêtes sur d'éventuels traces sémantiques (linguistiques et structurels) du besoin informationnel.

La deuxième partie est consacrée aux expérimentations. Avec tout d'abord dans le **chapitre 5** la présentation de nos différentes hypothèses puis la description de l'expérience réalisée pour les tester. Nous y présentons également plus en détails le corpus que nous avons réalisé à partir de demandes formulées en langue naturelle d'un service SAV d'un moteur de recherche, des profils utilisateurs, de leurs tâches à effectuer ainsi que les requêtes exécutées dans le moteur de recherche.

Enfin, la troisième partie présente les résultats obtenus ainsi que les interprétations qui peuvent en découler. Nous présentons dans le **chapitre 6** les traits caractéristiques des demandes en LN ainsi que les traits caractéristiques des requêtes. Nous établissons également des correspondances. Le **chapitre 7** apporte des précisions sur les résultats du chapitre 6 en les distinguant par typologie d'utilisateurs. Nous analysons alors de façon détaillée les besoins informationnels ayant le plus souffert ou bénéficié de ce passage de langue naturelle aux requêtes.

Des voies de recherches ultérieures découlant sont finalement présentées en conclusion.

Première partie

Etat de l’art sur les besoins informationnels et leur formulation via un système de recherche d’informations

CHAPITRE 1

LE BESOIN INFORMATIONNEL

La problématique centrale de ce travail est le lien entre le besoin informationnel d'une personne qui utilise un système informatique *i.e.* un utilisateur et sa formulation dans un Système de Recherche d'Informations (SRI).

Pour cela, nous relèverons en premier lieu les activités d'un individu lorsqu'il est confronté à un besoin informationnel ainsi que les processus inhérents relevés lors d'une tâche de RI.

1.1 Principes et définitions du besoin informationnel

Désigné comme « l'inconnu » des SRI par [SIM 06], le besoin informationnel résulte d'une prise de conscience par le sujet d'un manque de connaissances qui lui est nécessaire pour la résolution d'un problème ou pour l'atteinte d'un objectif visé et ce dans une situation donnée.

Le besoin informationnel serait donc la prise de conscience d'un utilisateur lorsqu'il est confronté à l'exigence d'une information qui lui est à la fois déficiente et nécessaire. Ce besoin apparaît comme étant ancré, déterminé par la position qu'occupe un individu dans son environnement social [LEC 98] ou de travail. Par exemple, des besoins émergent aux contacts d'autres personnes ou d'autres actions à la lecture d'un article ou lors de discussions avec des collègues de travail. Ces besoins peuvent également être révélés dans d'autres situations comme dans des activités de loisirs. Il s'agit d'un problème d'ordre cognitif : « c'est ce que je devrais avoir », « il me faudrait des information sur . . . ».

Dans le cadre des interviews de la liste Enseignant-Documentaliste sur le site « Doc pour docs », [TRI 04b] retrace les principales caractéristiques du besoin informationnel. Les principaux aspects en sont :

1. la prise de conscience d'un besoin d'informations,
2. la nécessité ressentie de combler une déficience d'informations,
3. le besoin de résoudre l'incertitude,
4. la contextualisation du besoin par rapport à des situations [TRI 03a].

L'auteur rassemble donc les différentes acceptations du besoin informationnel : à la fois mécanismes intellectuels et psychologiques face au questionnement, au doute, à l'incertitude et leurs étroites relations avec la connaissance.

[ING 96] distingue trois classes de besoins informationnels chez l'individu :

- **Besoin vérificatif** : l'utilisateur veut vérifier une information, retrouver des éléments d'informations qu'il possède déjà. Il sait même souvent comment accéder directement à l'information comme lors de la recherche d'un article sur Internet ou une date de sortie d'un ouvrage.

- **Besoin dont la thématique est connue** : l'utilisateur veut clarifier, revoir ou approfondir certains aspects d'un sujet déjà connu. Certaines notions relatives au sujet sont déjà acquises. Le besoin peut s'affiner et se préciser au cours de la recherche.

- **Besoin dont la thématique est inconnue** : l'utilisateur cherche des nouveaux concepts ou des nouvelles relations en dehors de sujets ou de domaines déjà familiers.

Lors des besoins *vérificatif* et de *thématique connue*, les demandes exprimées peuvent être déjà plus ou moins exhaustives selon les connaissances et le niveau d'expertise de l'individu. La demande peut être simple, par exemple quand un chercheur veut obtenir la date de parution d'un article. *A contrario*, le besoin de *thématique inconnue* est intrinsèquement varié et incomplet.

En d'autres termes, l'expression du besoin informationnel est l'énonciation de ce qui est inconnu à l'utilisateur [BEL 82a]. Cela peut paraître paradoxal ; l'utilisateur a suffisamment de connaissances pour savoir qu'il a besoin d'informations, mais pas forcément assez pour poser les bonnes questions qui lui fourniraient l'information pertinente. Pour [TRI 03a], pour prendre conscience d'un manque de connaissance, il faut déjà avoir des méta-connaissances.

1.2 Composantes conceptuelles et procédurales du besoin informationnel

Une fois que l'individu a conscience de son besoin informationnel, il va devoir conceptualiser cette demande avec des termes intelligibles par un SRI en se posant deux questions-clés :

1. Que vais-je chercher ?

2. Comment et où vais-je chercher ?

On distingue ici *composante conceptuelle* et *composante procédurale*.

- La **composante conceptuelle** (*i.e.* le but de la recherche) va s'élaborer et évoluer en même temps que les recherches. La représentation mentale du besoin informationnel peut être précise ou floue : les résultats vont servir à incrémenter et à enrichir voire à modifier le but initial de la recherche. La RI sera donc une suite d'actions liées entre elles dont chacune pourra potentiellement et individuellement influencer le cours de la recherche et le résultat final. [TRI 03a] indique également que le maintien en mémoire du but de la recherche est difficile car d'autres activités sont en cours comme la planification, la sélection, la compréhension, ou encore l'évaluation.

- La **composante procédurale** va obliger l'utilisateur à choisir un outil de recherche pour effectuer sa demande : systèmes documentaires informatisés, moteurs de recherche[1], outils d'interrogation de bases de données... pour ensuite traduire sa demande en *requête* interprétable par la machine.

Le recours à un SRI peut alors engendrer plusieurs problèmes selon les utilisateurs :

1. Ils peuvent ne pas savoir de quelles informations ils ont besoin, ou mal exprimer leurs demandes,

2. Ils peuvent aussi ne pas savoir où se trouvent les informations ou dans quelles sources les rechercher.

D'après [SIM 06], les SRI classiques ne sont pas du tout adaptés à la composante procédurale car l'utilisateur ne dispose pas souvent du vocabulaire adéquat pour formuler sa demande. Une traduction du besoin en une équation de recherche peut donc se révéler inappropriée selon les types de besoins informationnels et les profils utilisateurs.

Dans la même veine, [SHN 05] indique qu'il n'est pas toujours aisé de traduire son besoin informationnel en une requête interprétable par la machine car cela demande de faire concourir simultanément deux types d'opérations :

1. **Une opération cognitive** : on choisit ce que l'on doit demander et on le formule. Cette opération comprend donc la définition du sujet et du ou des énoncé(s) des idées principales ainsi que le choix des concepts,

[1] Un moteur de recherche est une application web permettant de retrouver des informations et des ressources associées à des mots, généralement appelés « mots-clés ».

2. **Une opération linguistique** : la question évolue d'une phrase intelligible en langage libre à une formulation dans laquelle on a supprimé des marqueurs sémantiques (les mots interrogatifs « Qu ») et dans laquelle on a introduit des signes méta-langagiers ainsi que des opérateurs de connexion, de troncature et de proximité.

1.3 Transformation du besoin informationnel en but de RI

De nombreuses modélisations de l'activité de transformation du besoin informationnel en but de RI ont vu le jour. Ces modèles présentent la particularité de décrire les processus cognitifs s'opérant chez les individus pendant les phases de recherches d'informations. Nous présentons ci-après les trois modèles qui rendent compte des différentes activités et processus intervenants lors d'une RI : le modèle de Guthrie, celui de Armbruster et Armstrong et enfin celui de Rouet et Tricot -lui même inspiré des deux premiers-.

1.3.0.1 Le modèle de Guthrie

[GUT 88] propose un modèle cognitif de l'activité de recherche d'informations en cinq étapes : la formation d'un but, la sélection d'un thème, l'extraction puis l'intégration de l'information, et enfin la modification *i.e.* la réitération des étapes précédentes jusqu'à l'obtention de l'objectif atteint.

1. **Formation du but** : l'individu doit être capable de se représenter un objectif à atteindre par rapport à une question donnée,

2. **Sélection d'un thème** : l'individu inspecte et choisit les différentes sources disponibles qui lui seront pertinentes pour la réalisation de sa tâche,

3. **Extraction de l'information** : l'individu extrait parmi les sources d'informations sélectionnées celles qui lui sont utiles pour atteindre son but ou répondre à sa question,

4. **Intégration** : l'information extraite est intégrée à l'information préalablement acquise. Elle sera combinée efficacement pour construire progressivement une synthèse,

5. **Modification** : l'individu réitère les quatre composantes du processus jusqu'à l'obtention d'une réponse globale satisfaisante.

Deux points sont perfectibles dans ce modèle : les critères d'identification, de compréhension et d'évaluation de l'information sont regroupés et réduits à la phase 3 (*i.e extraction*) et la modification et la réitération des étapes ne peuvent se produire qu'à la fin du processus [ROU 01].

1.3.0.2 Le modèle de Armbruster et Armstrong

L'étude de [ARM 93] porte principalement sur la définition des besoins et des activités de RI chez les enfants. Les auteurs identifient quatre éléments :

- **la formation du but** (*Goal formation*) : cet élément est identique au modèle précédent. Le processus commence par une identification de l'objet de la tâche de RI. Deux questions se posent :

  1. Quels types d'objectifs de recherche ces individus ont-ils ? Trois dimensions sont prises en compte : (i) les raisons de leur recherche peuvent être externes (*e.g.* questions demandées par leurs professeurs) ou internes (*e.g.* auto générées lors de leurs lectures), (ii) le moment où est apparu l'objectif de la recherche (avant ou pendant/après leur recherche), (iii) la spécificité de la tâche (spécifique, générale).
  2. Que savent-ils de la tâche de RI : peuvent-ils lancer leur propre tâche de RI ? Peuvent-ils traduire en une tâche de recherche appropriée, des objectifs qui leurs ont été imposés extérieurement ?

- **Sélection du texte** (*Text selection*) : après l'établissement d'objectifs, les individus sélectionnent le texte approprié pour répondre à leur recherche. Cette étape a été la plus délicate dans cette expérience : les individus éprouvaient des difficultés à assumer une « perspective développementale » de leur besoin. Elle a révélé notamment des traitements différents de la tâche entre les plus jeunes / les moins compétents et les plus âgés / plus compétents.

- **Extraction et intégration de l'information** (*Information extraction and integration*) : une fois le texte approprié choisi, les individus extraient et intègrent l'information pertinente à leur document. Plus le but de la recherche est spécifique,

moins les étapes d'extraction et d'intégration des informations sont susceptibles d'être exigées (dans une masse d'informations importante, la nécessité de faire un tri est primordiale). Les résultats suggèrent que les plus jeunes et les moins assidus en lecture sont les moins compétents dans les étapes d'extraction et d'intégration de l'information. Les meilleurs lecteurs et les lecteurs les plus âgés obtiennent de bons résultats particulièrement quand les recherches sont générales.

- **Evaluation** (*Evaluation*) : l'étape finale du processus est l'évaluation : elle détermine la réalisation (ou non) du but initial de recherche. Cette activité implique l'interrogation du lecteur sur ses stratégies d'apprentissage et met en rapport les moyens utilisés avec les résultats obtenus *i.e.* la métacognition. Elle permet également de stabiliser les procédures utilisées lors de ces différentes étapes. D'autres facteurs -autres que l'âge et les pratiques de lecture- affectent également les activités de recherche : la connaissance préalable de la structure d'un texte et le thème de la recherche.

L'étape *extraction et intégration de l'information* est une étape cruciale du présent modèle : elle intègre la complexité des processus cognitifs inhérents à la sélection et à la représentation des données contenues dans un document. Cette expérience révèle des pratiques différentes de RI et notamment d'auto-évaluation sur les procédures employées selon l'âge et la compétence des lecteurs.

1.3.0.3 Le modèle EST de Tricot et Rouet

A partir de ces modèles, [ROU 98] proposent d'examiner les processus cognitifs de RI dans les documents complexes *i.e.* dans lesquelles la réponse exige la consultation de plusieurs parties de textes. Ils font émerger deux caractéristiques de RI :

1. **sa nature cyclique** : le but d'une recherche est de satisfaire au mieux un ensemble de contraintes identifiées initialement par l'analyse de l'environnement. Généralement, l'activité de recherche ne s'arrête pas parce que le but est atteint mais parce que les résultats obtenus sont un compromis entre le temps passé, les moyens nécessaires, la motivation mise en jeu et la satisfaction des résultats obtenus ; l'utilisateur ne pense pas trouver mieux ou estime que le niveau de résolution est suffisant.

2. **son caractère hiérarchisé** : le *cycle de base* de toute RI est partiellement automatisable.

A partir de ces caractéristiques, les auteurs font l'hypothèse que toute RI comporte trois grandes phases : l'Evaluation, la Sélection et le Traitement (EST). Chaque phase se décompose en plusieurs processus plus spécifiques :

- **La phase Evaluation** : la RI commence par une représentation mentale de la tâche sans pour autant avoir les moyens de la réaliser : construction d'une représentation du but, comparaison des informations disponibles, productions de critères déclaratifs et procéduraux guidant la recherche.
- **La phase Sélection** : l'utilisateur décide d'examiner les catégories d'informations jugées intéressantes. Cette sélection peut engendrer une modification de ses critères initiaux.
- **La phase Traitement** : ensemble des processus se déroulant lorsque l'utilisateur examine une unité de contenu du système d'informations. Ces processus sont variables selon le type d'informations disponibles (textes, images, sons) et le contexte d'activités.

Ces modèles mettent en évidence plusieurs éléments : tout d'abord la RI est considéré comme un cycle de traitement constitué de trois phases principales : la sélection de l'information, le traitement de l'information sélectionnée et l'évaluation de la pertinence de cette information, en fonction du but visé par le sujet. Deuxièmement, dans le modèle EST, l'activité de sélection et d'évaluation recouvre un processus de gestion de l'activité (planification de la recherche et évaluation de l'écart entre la situation actuelle et le but visé) et un processus de traitement des informations relationnelles (liens, menus, boutons). Enfin, ces processus se conduisent en relation avec la représentation que l'utilisateur se fait de la tâche. Celle-ci inclut une représentation du but et peut être modifiée dynamiquement au cours de l'activité de recherche.

Une recherche d'informations réussie est donc être le fruit d'une prise de conscience d'un besoin informationnel, d'un raisonnement organisé dont les objectifs sont présents tout au long du processus et d'une intériorisation des stratégies de recherche. Dans la section suivante, nous abordons plus longuement la conceptualisation du besoin informationnel au sein même d'un SRI.

CHAPITRE 2

PRISE EN COMPTE DU BESOIN INFORMATIONNEL À TRAVERS L'UTILISATEUR-ACTEUR DU SRI

Les SRI qui émergent entre les années 50 et 60 sont assez rudimentaires : les données contenues dans les premières bases sont initialement fournies sous forme de listings, sans aucun tri, filtre, ou sélection possible. La création d'index permet d'effectuer les premières requêtes exclusivement lancées par des documentalistes qui traitent et répondent des demandes des utilisateurs formulées sous forme papier : elles doivent alors utiliser et maîtriser des langages de requêtes spécifiques pour consulter les bases de données.

La situation dans la fin des années 60 évolue considérablement avec la possibilité de se connecter à un ordinateur distant : les requêtes s'effectuent sur des terminaux affectés à certaines bases pour la consultation. Les recherches et expérimentations sont principalement centrées sur le système : traitements de l'information et représentations des documents dans un SRI. C'est ce que [ELL 92] désignera plus tard comme étant le **paradigme physique** *(physical paradigm)*. On trouve également les désignations suivantes : approche traditionnelle par [ING 92], paradigme ou approche orienté(e) système (*system oriented*), paradigme système [CHA 04a] ou encore paradigme classique orienté système [POL 00]. Il est généralement opposé au **paradigme cognitif** *(cognitive paradigm)*. Nous détaillons dans la section suivante comment nous sommes passés de l'approche orientée systèmes où la prise en compte de l'utilisateur était quasi inexistante à une approche orientée « client » ou « usager » où l'utilisateur devient primordial. En effet, l'utilisateur est alors placé au centre des préoccupations, pour devenir un des *acteur* principal du SRI : on ne s'intéresse plus simplement à sa requête mais aussi à son profil, ses compétences, ses connaissances, le contexte de sa recherche.

2.1 Du SRI orienté « systèmes »

Les propriétés de cette approche portent principalement sur les aspects techniques du système de recherche :

- *le traitement de l'information* : extraction de l'information dans les textes sous forme de concepts et de termes pour la gestion, le contrôle, l'indexation et la représentation des documents,

- *les possibilités d'interrogation* : les règles syntaxiques, les termes de commande, les opérateurs définis pour la formulation des requêtes et le repérage de l'information dans les documents,

- *les algorithmes de correspondance et d'appariement* : entre les termes d'indexation et les termes de la requête.

L'objectif est alors l'amélioration de la performance intrinsèque du SRI ; l'évaluation du système devient une préoccupation majeure notamment avec la mise en place de métriques de comparaison.

2.1.1 Caractéristiques de l'approche orientée système

Pour [SAR 96] les représentations du modèle orienté système de RI sont assimilées à deux ensembles : *(1)* le système et *(2)* l'utilisateur dont les éléments et les processus convergent vers un filtrage et une mise en correspondance des résultats.

La représentation de ce système, illustré par la Figure 2.1, implique que les objets informationnels (textes, images, données...), aient été représentés et organisés dans des dossiers afin qu'ils puissent être exploités au moment du filtrage et de la mise en correspondance. La représentation de l'utilisateur débute avec l'apparition d'un problème ou d'un besoin informationnel ; représenté le plus souvent par une question et transformée en une requête acceptable et compréhensible par la machine. A l'intersection de ces deux représentations (documents et requêtes), une mise en en correspondance et une comparaison s'opèrent.

Cette mise en correspondance permet ainsi de retrouver des documents au degré de pertinence calculable. Cela permet aussi d'évaluer les performances d'un SRI en fonction de deux mesures : le rappel (équation 1), la précision (équation 2). Le rappel est la proportion de documents pertinents renvoyés par le système parmi tous ceux qui sont pertinents dans une collection. La précision est la proportion des documents pertinents parmi l'ensemble de ceux renvoyés par le système.

$$rappel = \frac{|documents\ pertinents \cap documents\ retrouvés|}{|documents\ pertinents|} \quad (1)$$

$$précision = \frac{|documents\ pertinents \cap documents\ retrouvés|}{|documents\ retrouvés|} \quad (2)$$

Dans cette approche, un système capable de trouver l'ensemble des documents per-

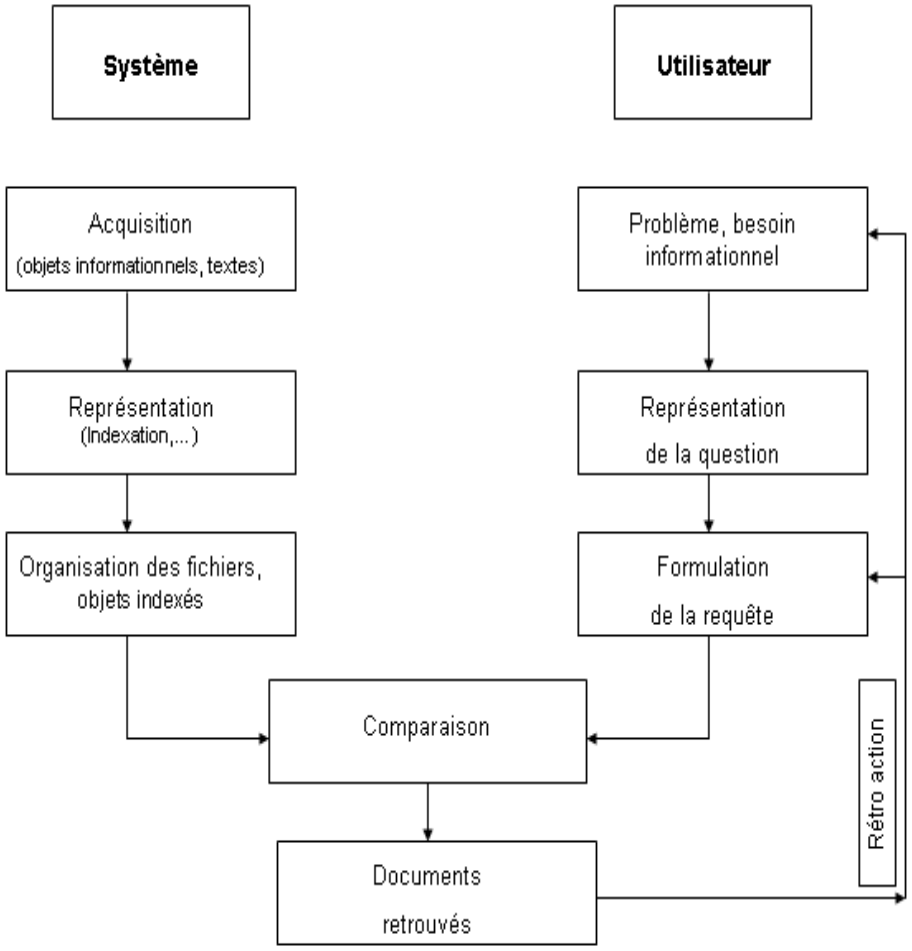

Figure 2.1 – Modèle de RI traditionnelle selon Saracevic

tinents et en éliminant dans sa sélection le maximum de documents non pertinents est jugé performant. L'utilisateur est appelé à considérer de la pertinence des documents qu'en deuxième instance. A noter qu'une notion de rétro-action (*feedback*) permet une modification des représentations soit au moment de l'identification des besoins, soit au moment de la saisie des requêtes.

2.1.2 Analyses et critiques de l'approche orienté système et de son évaluation

Plusieurs expérimentations vont voir le jour à partir des années soixante dont l'objectif principal est d'évaluer les compétences et les fonctionnalités du SRI et de les tester expérimentalement. Parmi ces expérimentations, nous pouvons citer : le projet Cranfield (dirigé par Cleverdon, 1957-1967), le projet Smart (dirigé par Salton, 1964), le projet Medlars (dirigé par Lancaster, 1966-1967), le projet TREC (dirigé à ses débuts par Harman, à partir de 1992), le projet stairs (dirigé par Blair et Maron, 1985). Les principales critiques de [IHA 99], [CHA 04a], [CHA 04b] de cette approche sont centrées sur l'incapacité à rendre compte des paramètres qui peuvent rentrer en compte lors d'une RI dans un système, à savoir :

- **Le besoin informationnel est considéré comme statique** : [BEL 93] juge cette approche -et notamment la représentation standard de la RI en deux ensembles (système et utilisateur)- inadéquate car elle impliquerait que le besoin d'informations soit statique, c'est à dire que l'utilisateur du système ait une représentation déjà façonnée de son besoin et de sa représentation sans prise en compte d'une éventuelle évolution lors de ses recherches. Or, comme nous l'avons vu dans la section 1, le besoin informationnel évolue tout comme les connaissances et l'expertise des utilisateurs. Le SRI doit donc en tenir compte.

- **L'interaction entre le système et l'utilisateur est mal représentée** : le système dévalue l'importance de l'interaction entre le système et l'utilisateur en ne fournissant qu'une seule forme de comportement. Pour [SAR 96], l'interaction n'est pas directement représentée mais seulement supposée et englobée dans la rétroaction (*feedback*) ; ce qui est insuffisant pour prendre en compte l'évolution des recherches.

- **Les critères de jugement de pertinence des documents sont inadéquats** : avec les méthodes utilisées, les documents qui ne font pas partie de l'appariement système-utilisateur ne sont pas jugés pertinents. La pertinence est alors représentée comme une décision binaire (pertinent, non pertinent) sans aucune nuance ni granularité alors que les documents peuvent être partiellement pertinents, ne portant que sur une partie du texte par exemple. [SCH 94] a identifié 80 facteurs agissant et/ou influençant sur la performance d'un document. Ces variables sont donc à prendre en considération pour les valeurs de rappel et de précision qui pourraient

être faussées et dont [TAG 92] remet en cause pour l'évaluation de la pertinence des documents.

2.2 Au SRI orienté « end-user »

A partir des années 80 et en regard des remarques et des lacunes adressées à l'approche orientée système, se dresse une nouvelle approche qui prend en compte le comportement réel des utilisateurs dans l'évaluation des systèmes. En effet, les besoins et les usages variés des utilisateurs sont étudiés afin d'être mieux représentés lors de l'interaction homme-machine (IHM).

2.2.1 Caractéristiques de l'approche orientée « end-user »

Dans cette approche, l'interaction devient le dispositif le plus important de la RI. Pour [POL 00], les systèmes ont pour objectif de faciliter la communication entre un producteur d'informations (l'auteur) et un utilisateur.

Cette approche rassemble plusieurs travaux dont notamment :

- le modèle ISP (*Information Search Process*) de [KUH 91] : Kuhlthau considère que le processus de RI est affecté par des états émotionnels des utilisateurs du système ; il peut s'agir de l'incertitude, la confusion ou encore l'anxiété.

- le modèle « en épisodes » de [BEL 82a] [BEL 85] : Belkin met en évidence les changements d'états de connaissance dans la RI. Il part du principe que le point de départ de la recherche se caractérise par un état de connaissance « anormal » ou « déformé » *Anomalous State of Knowledge* qui va se prolonger en une suite de différentes interactions possibles : jugement, interprétation, modification, navigation où la temporalité joue un rôle primordial. En effet, l'utilisateur s'engage dans un nombre et un temps fini d'interactions ; cette temporalité influençant la nature des interactions.

- le modèle cognitif de [ING 92] [ING 96] : Ingwersen a pour objectif d'identifier les processus cognitifs et les différentes éléments impliqués qui interviennent dans le processus de RI. Ainsi, les intentions, les expériences, les influences, les préférences, les connaissances sociales et individuelles sont ainsi reflétés dans le processus de RI. Les interactions complexes entre les éléments impliqués dans le processus global de RI sont indiquées dans son modèle [ING 92] : les objets

informationnels, la configuration du SRI, l'espace cognitif de l'utilisateur, l'environnement social et organisationnel.

- la formalisation du comportement global de l'utilisateur selon [WIL 96] : Wilson s'appuie sur trois éléments pour proposer un modèle général du comportement informationnel : *(1)* le besoin informationnel et ses origines *i.e* les facteurs qui conduisent à la perception du besoin par l'individu, *(2)* les facteurs qui déterminent la réponse de l'individu en réaction à la perception du besoin, *(3)* les processus ou les actions qui sont impliqué(e)s par cette réponse. Wilson indique qu'il existe des facteurs qui facilitent ou au contraire freinent l'accès à l'information. Il introduit en effet plusieurs théories, en s'appuyant sur d'autres disciplines pour modéliser le processus global de RI. Parmi ces théories, notons celles du stress (*stress/coping theory*), du risque (*risk/reward theory*) ou encore celle de l'auto-efficacité (*social learning theory* avec la notion de *self-efficacy*). Ces théories tentent d'expliquer les stratégies de recherche individuelle : pourquoi certaines sources d'informations sont utilisées et d'autres non, pourquoi certains besoins n'impliquent pas nécessairement la mise en place d'une stratégie de recherche effective ou encore comment l'utilisateur adapte son comportement à celui requis afin d'arriver au résultat escompté.

2.2.2 Analyses et critiques de l'approche orientée « end-users »

Ces modèles ont pour point commun de proposer une modélisation des utilisateurs et de leurs comportements. Pour [CHA 02], cette approche produit des schémas qui sont encore trop peu exploitables pour l'évaluation des SRI. Ces deux auteurs soulignent notamment l'importance de lier l'évaluation des systèmes à l'analyse des besoins informationnels. En effet, les besoins évoluent durant les différentes interactions avec le système, il n'est donc plus possible de les considérer comme définis *a priori*. Ces besoins doivent au contraire être appréhendés en amont du cycle de développement des systèmes puis constamment réactualisés durant tout le cycle de recherche. Il est également indispensable de re-contextualiser le besoin informationnel : *i.e* le reconstruire ou le remodeler *a posteriori* en fonction du type d'utilisateur, ses objectifs, sa tâche à effectuer, ses exigences… afin de répondre à la difficulté d'identifier l'utilisateur. L'ensemble des informations et indices recueillis en amont ou au cours des interactions doivent être mieux analysés et appréhendés par le SRI.

Ce modèle se focalise donc davantage sur la notion d'utilisateur et sur une meilleur

prise en compte du besoin informationnel. La réflexion menée sur cette approche nous permet d'envisager un nouveau paradigme plus complexe, plus précis prenant en compte la contextualisation de l'utilisateur-acteur dans sa tâche à effectuer.

2.3 Vers un SRI contextualisé orienté « tâches »

La modélisation de l'utilisateur -et notamment de sa tâche à accomplir- est devenue essentielle dans les champs de l'IHM et dans le même temps difficile à mettre en place au sein d'un SRI. En effet, les SRI actuels sont souvent trop généralistes ; mettant en œuvre les mêmes mécanismes et les mêmes méthodes de traitement de l'information, et ce, quel que soit le contexte de recherche, l'utilisateur, son type de besoin informationnel et l'usage qu'il souhaite faire de l'information retrouvée [MOT 11]. L'approche trop généraliste des outils disponibles de RI sur le web est à l'origine des problèmes de dégradation de la qualité des résultats retournée par les SRI [BUD 00]. C'est en prenant en compte dans un premier temps les connaissances explicites de l'utilisateur mais aussi celles implicites que les nouveaux modèles de RI pourront évoluer et s'améliorer. En effet, les SRI devront savoir prendre en charge lors des différentes interactions avec l'utilisateur les éléments non explicites, *i.e.* non verbalisés par l'utilisateur, à savoir : ses intentions, son environnement, son expertise, son expérience… tous les éléments susceptibles d'améliorer les connaissances sur les utilisateurs. Il s'agit alors pour les SRI de construire un modèle de l'utilisateur basé sur deux types de connaissances :

- celles connues *a priori* sur lesquelles le système construit le profil utilisateur : les paramètres entrés lors d'une première recherche sont pris en compte. Notons ici le problème du *« démarrage à froid »* relevé [MEY 11] (*i.e.* aucune connaissance sur l'utilisateur lors de sa première recherche) qui conduit à des performances très pauvres pour les premières recherches des nouveaux utilisateurs,

- celles issues des connaissances construites de façon dynamique au fur et à mesure des interactions avec le système avec l'historique des recherches, les reformulations, le type et le nombre de documents consultés.

[MOT 11] rappelle qu'un SRI a pour but de satisfaire des besoins d'informations, et que cette satisfaction dépend de l'objectif de l'utilisateur et de l'usage qu'il souhaite faire de l'information ; selon qu'il s'agisse de vérifier une hypothèse, répondre à une question précise, réaliser un rapport. Les réponses à fournir à l'utilisateur doivent prendre en compte ces caractéristiques : elles peuvent être concises (une phrase, un paragraphe) tout

en répondant à une question précise, alors que pour une étude, un plus grand nombre de documents sera attendu. La redondance dans les réponses peut correspondre à un besoin de vérification d'une information ou au contraire à un bruit, en fonction de la recherche. L'unité d'informations peut donc varier en fonction du besoin et de la **tâche à réaliser**, d'où l'importance de se focaliser sur ce dernier aspect.

2.3.1 Caractéristiques de l'approche contextualisée orientée « tâches »

Le contexte couvre des aspects assez larges tels que l'environnement cognitif, social et professionnel dans lesquels s'inscrivent des situations liées à des facteurs tels que le lieu, le temps et l'application en cours. Les auteurs qui se sont penchés sur cette question ne convergent pas tous vers une même définition, même si on retrouve des dimensions de descriptions communes comme l'environnement cognitif, le besoin mental en information et l'interaction liée à la recherche d'informations [COO 02].

Des travaux en recherche contextuelle d'informations (RCI) ont vu le jour ces dernières années. L'objectif est d'optimiser la pertinence des résultats de recherche en impliquant *(1)* la définition du contexte du besoin informationnel de l'utilisateur puis *(2)* en l'adaptant à la recherche et en le prenant en considération dans le processus de sélection de l'information.

Parmi les éléments contextuels les plus importants traités dans la littérature, nous pouvons citer le profil de l'utilisateur, le contexte de la requête, le contexte d'interactions avec le système, le temps de soumission de la requête, les préférences de recherche liées à la qualité de l'information (fraîcheur de l'information, genre du document, crédibilité de la source de l'information, etc.), le contexte temporel et géographique de recherche.

A noter que peu de travaux en RCI ont exploré la dimension du contexte liée aux préférences de qualité de l'information, sa fraîcheur [LI 03] [ROD 06] ou encore la crédibilité de la source [HAR 06]. Plus récemment, des travaux ont vu le jour sur le contexte géographique de l'utilisateur et plus spécifiquement sur les utilisateurs d'appareils mobiles [BOU 09c]. L'idée de base est d'apprendre les centres d'intérêt de l'utilisateur pour chaque situation donnée : une situation correspondant à l'agrégation des contextes spatio-temporels issus des activités de recherche passées de l'utilisateur. Les centres d'intérêt de l'utilisateur sont donc appris à partir des activités de recherche passées relatives aux situations ainsi identifiées. Des études sur sur les *logs* des requêtes des utilisateurs mobiles [KAM 07] montrent que les requêtes sont plus courtes et qu'il y a moins de requêtes par session. D'après [SOH 08], 72 % des besoins informationnels

des utilisateurs mobiles sont liés à des facteurs contextuels comme la localisation et le temps.

Plusieurs taxonomies du contexte en RI (illustrées par la Figure 2.2) ont été proposées parmi lesquels celles de [GOK 08], [TAM 10] et [TAM 03] qui reprennent certaines dimensions spécifiques du contexte, à savoir :

- **le dispositif technique** : outil physique qui permet à l'utilisateur d'accéder à l'information,
- **la tâche/ le problème** : intention ou but de l'activité de recherche,
- **le contexte du document** : caractéristiques des sources d'informations et métadonnées des documents (formes, éléments de structure, citations,etc),
- **le cadre spatio-temporel** : dimensions de temps et de localisation géographique,
- **le contexte utilisateur** : lié au contexte social de l'individu et au contexte personnel (démographique, psychologique, cognitif). Le contexte démographique regroupe les attributs de préférences personnelles comme la langue ou le genre [HUP 06]. Le contexte psychologique regroupe les attributs comme l'anxiété et la frustration [KIM 08]. Le contexte cognitif réfère au niveau d'expertise de l'utilisateur [TIM 05] et à ses centres d'intérêt à court terme [DAO 09b] ou à long terme [TAM 08].

Pour [CAL 06], le contexte de la RI dépend de la date, du lieu, de l'historique de l'interaction, tâche en cours et d'autres facteurs non explicités mais implicites dans l'interaction et l'environnement de la recherche (Figure 2.3). La contextualisation de la RI doit également faire face à l'émergence de nouveaux environnements de recherche d'information comme la téléphonie mobile [BOU 09c], les agendas personnels pour lesquels le contexte constitue une composante important.

Plus simplement, [CHA 10] propose une modélisation du contexte en RI reposant sur un triptyque utilisateur-tâche-environnement :

1. **l'utilisateur** : ses centres d'intérêts à court et long termes, ses habitudes, sa catégorie socio-professionnelle, ses différentes expertises, ...
2. **la ou les tâche(s) à réaliser** : il est intéressant de définir ici si les tâches à effectuer s'effectuent dans un environnement professionnel ou personnel / de loisirs,

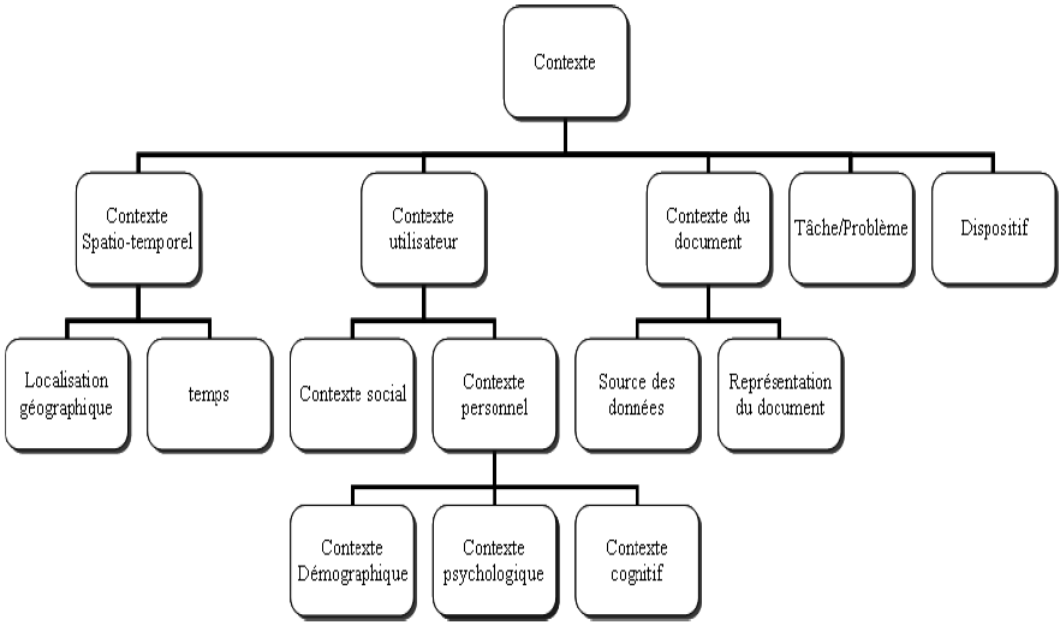

Figure 2.2 – Taxonomie du contexte en RI selon [TAM 10]

3. **l'environnement de la recherche** : peut concerner la localisation géographique mais aussi plus techniquement le matériel utilisé comme le logiciel ainsi que le réseau, la bande passante, etc. [STO 05], [BRU 01]

Nous décrivons plus longuement dans la section suivante ces trois parties interdépendantes.

2.3.2 La contextualisation de l'utilisateur

[CAB 11] font part d'une réflexion sur la place de l'utilisateur dans le développement des SRI. À la base, pour tous, un utilisateur est un individu qui, dans un contexte donné -professionnel ou personnel- a besoin des résultats de sa requête médiatisée par un système informatisé -un logiciel quelconque, ou un système de recherche d'informations- pour réaliser une tâche avec un objectif spécifique.

Cette réflexion, menée en parallèle par des membres des communautés informatique et ergonomie cognitive, peut apporter des éclairages spécifiques et complémentaires. Ces communautés appréhendent en effet l'utilisateur différemment.

- La communauté informatique s'attache à concevoir des logiciels génériques qui

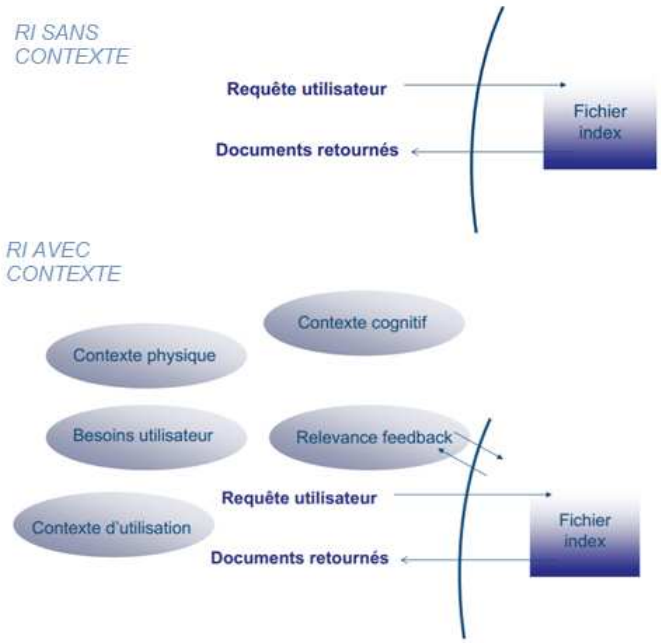

Figure 2.3 – Contexte en RI selon [CAL 06]

pourront ensuite s'adapter à l'utilisateur afin d'être personnalisés - démarche qualifiée d'*ascendante*. Cette démarche part des fonctionnalités que l'on souhaite implémenter dans le système logiciel en observant ce que fait l'utilisateur (traces) comme par exemple le temps de lecture comme indicateur de la perception humaine de la pertinence.

- La communauté d'ergonomie cognitive observe l'utilisateur pour en déduire les spécifications d'un système informatique adéquat qui lui conviendrait mieux. Cette démarche est qualifiée de *descendante* ; partant de l'analyse de l'utilisateur et de son comportement pour identifier des éléments pouvant être exploités dans les systèmes comme par exemple l'impact de l'état émotionnel sur la réalisation de la tâche.

Ces deux approches convoitent toutefois le même objectif : concevoir des systèmes susceptibles d'aider l'utilisateur à réaliser au mieux sa tâche. Mais concevoir un tel système revient à être performant sur deux principes :

1. le premier étant de connaitre l'utilisateur : Qui est t-il ? Où se trouve t-il ? Que veut-il faire ou que cherche t-il ? Ces connaissances sur l'utilisateur seront à défi-

nir en fonction de chaque particularité des SRI : il n'existe pas de modèle d'utilisateur de référence, chaque application définissant et utilisant son propre modèle en adéquation avec les besoins spécifiques sous-jacents ;

2. le second étant de construire un « profil », appelé modèle utilisateur, rassemblant les connaissances issues de la première phase pour ainsi adapter le SRI en fonction de ce profil : fonctionnalités de recherche, présentation des résultats, unités documentaires sélectionnées.... Les connaissances sur le profil utilisateur pourront évoluer le cas échéant au cours de l'interaction avec le système.

Parmi les connaissances qui permettront d'établir le modèle utilisateur, nous pouvons relever :

1. **Les données personnelles et ré-actualisables de l'utilisateur** : elles fournissent directement des informations sur son identité, sa profession, ses centres d'intérêt.... Pour [BRU 01], les connaissances sur l'utilisateur peuvent concerner le sexe, l'age, le statut professionnel mais aussi relever d'autres types d'informations comme le niveau d'expertise du domaine ou encore les centres d'intérêts. Les connaissances individuelles peuvent être réactualisées au fur et à mesure des interactions et des échanges. Dans [BRE 04], la détection du niveau de familiarité de l'utilisateur avec un service permet d'adapter le service aux attentes et aux capacités de l'utilisateur.

   Dans les modèles de dialogue homme-machine, beaucoup de connaissances sont acquises durant les phases de dialogue. Certaines connaissances sont issues directement des dialogues (*Si vous êtes un expert, dites...*), alors que d'autres informations sont issues de certaines observations du système [SHI 92] comme la durée, les mots et concepts employés, les actions effectuées.

2. **l'historique et l'enrichissement des requêtes** : les historiques mémorisent la chronologie des besoins en informations de l'utilisateur. Les documents consultés lors de la recherche peuvent donner des informations supplémentaires pour construire le profil utilisateur. Certaines approches permettent d'enrichir ou de raffiner la requête initiale des utilisateurs, notamment dans la thèse de [DAO 09] qui utilisent des ontologies[1] pour recontextualier des mots-clés dans un domaine

[1]Une ontologie est un ensemble structuré des termes et concepts représentant en soi un modèle de données représentatif d'un domaine.

spécifique. Si nous prenons l'exemple du domaine d'intérêt de l'utilisateur, le résultat de recherche du terme *« bus services in Java »* ne doit pas être le même pour un informaticien et un touriste voulant aller sur l'île de Java en Indonésie. Le problème actuel étant que le SRI prend les mots-clés *« bus » « services » « Java »* pour déterminer des résultats et les présenter à l'utilisateur. En intégrant le domaine de l'utilisateur, le SRI pourra ajouter les termes *« langage de programmation »* à la requête initiale de l'informaticien ou *« tourisme »* dans le second cas [BAI 08]. Le contexte du domaine permet donc d'intégrer un ensemble de termes de base qui sont souvent omis par les utilisateurs eux-mêmes durant leurs phases de recherches.

3. **l'historique de navigation** : il mémorise la trace des pages visitées par l'utilisateur pendant sa recherche . Ces traces sont enregistrées à partir d'un serveur proxy et des *logs* de requêtes, permettant ainsi d'observer et de dessiner des groupes d'utilisateurs puis faire des suggestions et des recommandations.

4. **des jugements de pertinence** : ils consistent à donner des préférences sur les documents préalablement sélectionnés par le système. Ils permettent aussi de réinjecter des préférences voire de reformuler la requête initiale de l'utilisateur.

5. **les signets *(bookmarks ou folders*)** : ils mémorisent les URL des documents que l'utilisateur souhaite conserver. On peut également relever dans ce point l'organisation hiérarchique des ressources conservées par l'utilisateur ; sa façon d'organiser les informations peut également refléter explicitement ses centres d'intérêts [JON 07], [JON 05].

6. **les tags** : ce sont des marqueurs sémantiques ou lexicaux utilisés sur les sites dits de réseaux sociaux Web 2.0. L'intérêt de les étudier est qu'ils sont choisis librement par l'utilisateur pour décrire les ressources qu'il souhaite partager *via* un service social comme le bookmarking par exemple ou plus rarement pour des ressources qu'il désire tout simplement conserver. Les tags peuvent être de bons indicateurs quant aux centres d'intérêts de l'utilisateur [CAR 08].

7. **le réseau social de l'utilisateur** : le principe étant le partage social pour faciliter l'accès à l'information *i.e.* bénéficier des recommandations issues des autres recherches effectuées par d'autres utilisateurs [KIR 08].

Ces connaissances sur l'utilisateur servent ensuite à prédire le comportement [LIT 02] voire à l'anticiper dans certains cas [BAU 08].

2.3.3 La contextualisation de la tâche à réaliser

Une tâche est vue comme un ensemble d'actions (physiques, affectives et/ou cognitives) dans la poursuite d'un certain but qui peut être modifiable, évolutif. Cette définition suppose qu'une tâche ait -du moins quand elle est réalisée-, un début et une fin reconnaissables et identifiables. Elle indique également qu'une tâche a un objectif réalisable (*i.e. le résultat*) et un objectif significatif (*i.e. la raison*). Or, ces deux façons de définir les tâches sont toutefois assujetties à des points de vue subjectifs ou objectifs. Les tâches objectives sont externes à l'utilisateur. Ce point de vue est généralement adopté dans la recherche qui utilise des descriptions des tâches. *A contrario,* les tâches subjectives sont celles qui sont internes à l'utilisateur ; subordonnées à sa compréhension.

Par ailleurs, une tâche peut consister en de plus petites sous-tâches spécifiables avec chacune des conditions et des objectifs individuels simples. Ces sous-tâches ont besoin d'être appréhendées ensembles pour former un tout cohérent et significatif.

2.3.4 L'environnement de la recherche

L'environnement est indéniablement inter-connecté aux contextes et tâches utilisateurs. Les tâches n'existent pas indépendamment mais en interaction avec le contexte dont elles font partie [TAY 91], [ELL 97]. En effet, lorsque nous étudions les tâches, il est nécessaire voire recommandé de connaître en plus des facteurs individuels, les facteurs contextuels ainsi que ceux liés à une situation -[HAC 98], cité par [BYS05]- comme les tâches, les descriptions de tâches et le processus de l'exécution des tâches. C'est pourquoi, les facteurs contextuels et situationnels doivent être identifiés et pris en compte afin de construire une image globale de la façon dont les tâches sont interprétées. La conjonction de tous ces facteurs réunis crée la situation particulière qui sera le point d'entrée du SRI [CHA 10].

2.3.5 Analyses et critiques de l'approche orientée « tâches »

Pour résumer, un SRI peut être amélioré en *(1)* modélisant, *(2)* intégrant, *(3)* exploitant le contexte [CHA 10]. Toutefois, [ALL 02] indique que malgré un intérêt récent pour ces problématiques, peu de progrès ont été faits dans ce sens. Ceci probablement

à cause de la difficulté à extraire et à représenter les connaissances sur les utilisateurs, son contexte et ses tâches. Les améliorations réellement constatées du point de vue des utilisateurs sont assez rares. Lorsque les tâches sont considérées comme des processus, l'objectif est de comprendre les différentes actions qui ont lieu au cours d'une exécution de la tâche. Les chercheurs essaient de reconnaître comment les individus perçoivent leur tâche et de comprendre pourquoi et comment les différentes informations et leurs sources sont utilisées lors de leurs exécutions. Les résultats permettent d'augmenter la compréhension de l'information liée à l'action et conduisent souvent à des recommandations sur la façon de rechercher. Une des critiques que l'on pourrait désigner est le fait que ces résultats sont difficiles à généraliser car il y a un manque de conditions expérimentales sur lesquelles les fonctionnalités du système pourraient être évaluées. Cette évaluation prendrait en compte des comportements différents et des séquences de comportements dans lesquelles des variables indépendantes et des caractéristiques situationnelles pourraient être manipulées et étudiées.

Une évolution possible des SRI serait la capacité d'utiliser le contexte et les caractéristiques des requêtes pour inférer des réponses spécifiques et spécialisées aux utilisateurs *i.e*. des réponses différenciées selon le type de la demande, le niveau d'expertise de l'utilisateur, sa tâche à effectuer et pour cela adapter la ou les réponse(s) à fournir : des phrases, des passages de texte, des graphiques, des documents voire des combinaisons de paragraphes par exemple.

Nous étudierons dans la section suivante plus spécifiquement les différentes formes d'expression des besoins informationnels au travers d'un SRI : exploration thématique, langage de requêtes, langage naturel et les spécificités et adaptations que nous proposons pour une meilleure adéquation au système.

CHAPITRE 3

LES DIFFÉRENTES FORMES D'EXPRESSION DES BESOINS INFORMATIONNELS AU TRAVERS D'UN SRI

Un besoin informationnel peut être exprimé au travers d'un SRI sous différentes formes selon qu'il s'agisse principalement de logiciels documentaires ou de moteurs de recherche mais aussi de catalogues en lignes ou de portails.

Traditionnellement, les logiciels documentaires ont pour objet principal la gestion du fonds documentaire ; les champs permettant d'accéder à l'information sont alors paramétrés en fonction de cet objectif (noms d'auteurs, dates, etc.). Le mode de recherche de ces logiciels est souvent booléen (*i.e* équation logique entre les termes de la requête) avec la possibilité de rechercher sur la totalité du texte avec le mode *full text* sur certains champs comme des articles, des textes de loi ou de la documentation technique. L'expression du besoin informationnel se fait donc principalement par la saisie de mots-clés dans les différents champs proposés.

En ce qui concerne les moteurs de recherche, les modes de recherche proposés sont : l'exploration thématique, le langage de requêtes et le langage naturel. Ces modes de recherche ne sont pas exclusifs ; certains moteurs proposent de les combiner pour une meilleure utilisation de leurs services. Ces utilisateurs peuvent faire :

- de l'exploration thématique ; détaillée dans la partie 3.1,
- de la saisie de mots-clés ; présentée dans la partie 3.2,
- ou encore utiliser des interfaces permettant l'expression en LN ; explicitée dans la partie 3.3.

3.1 Traduction du besoin informationnel par une exploration thématique

Certains moteurs de recherche et annuaires[1] proposent un recensement des sites web classés par catégories et sous-catégories. Cet accès à l'information permet à l'utilisateur d'accéder directement à des thématiques sans avoir à sélectionner des mots-clés. Cette navigation peut se faire soit de façon arborescente, soit de façon hypertextuelle, soit de façon cartograhique (*topic maps*). Nous les explicitons dans la partie suivante.

[1] Les annuaires sont des répertoires de sites validés humainement.

3.1.1 La recherche par navigation arborescente

La recherche par navigation arborescente correspond à une démarche systématique où l'information est structurée et organisée selon des structures hiérarchiques : l'utilisateur peut, à partir d'un thème général, aboutir à des informations spécifiques. Ce mode d'accès permet tout d'abord de prendre connaissance de grandes thématiques à partir des menus ou des dossiers proposés. La navigation se présente ensuite comme un chemin à parcourir à travers des divisions ou rubriques puis des subdivisions ou sous-rubriques jusqu'à l'obtention de l'information désirée. On retrouve la navigation arborescente dans les plans de classement ou les annuaires thématiques sur Internet *e.g. Yahoo, Nomade*... Ce mode facilite le recouvrement général d'un thème et offre donc un meilleur rappel (voir définition page 10) que les modules d'interrogation.

3.1.2 La recherche par navigation hypertextuelle

La navigation hypertextuelle est un système de navigation dans un document ou entre documents, qui repose sur le principe de l'association, sur le chaînage d'une notion à une autre et qui fonctionne en définissant un réseau de nœuds et de liens entre ces nœuds d'informations. Les liens hypertextes permettent de relier des pages web, à partir des éléments textuels ou autres comme les images par exemple qui supportent ces liens. Cette recherche correspond à un profil à dominante lien hypertexte : pour parvenir à l'information cherchée, les utilisateurs suivent les liens proposés par l'interface. La démarche est associative : l'utilisateur parcourt et accède aux informations recherchées en suivant les liens proposés et par associations de mots et/ou documents. Cette liaison s'effectue entre les objets de recherche et les libellés ou étiquettes -également appelés « *tags* »- d'identification des contenus des objets informationnels.

Tout comme la recherche par navigation arborescente, l'utilisateur n'a pas à formuler son besoin par la saisie de mots-clés dans un formulaire de recherche. Ce type de recherche ne nécessite que très peu voire pas de connaissance préalable et permet un élargissement du thème de recherche. Pour autant, c'est un mode de recherche dont la navigation peut se révéler intempestive ; il est possible de se noyer dans une masse d'informations trop importante. Pour [ROU 98], l'utilisation d'un système d'hypertexte suppose que le lecteur soit en mesure de gérer son propre parcours dans l'information.

3.1.3 La recherche par interface cartographique

La recherche par interface cartographique -également appelé *topic map*- connait un certain essor depuis quelques années. Le principe de ce type de recherche est que chaque utilisateur peut-être associé à des synonymes, et un même terme peut être associé à plusieurs concepts. L'assemblage de plusieurs *topic* permet ainsi de former un graphe sémantique, contextualisant les voisins sémantiques proches du ou des termes recherché(s). Le principal atout de cet outil est de proposer une navigation plus pédagogique (créative et intuitive), et une vision d'ensemble du thème principal de la recherche et des concepts associés. Parmi les moteurs offrant ce mode de recherche, *KartOO* (fermé depuis janvier 2010) fut un précurseur dans le domaine de l'illustration cartographique des résultats d'une recherche. D'autres moteurs de recherches par interface cartagrophique ont vu le jour comme notamment *Serelex*[2]. Ce moteur propose des résultats sous deux formes :

- soit sous une forme de liste de termes co-ocurrents le plus avec le terme recherché,
- soit sous une forme cartographique dont nous donnons un exemple autour du terme *Banana* à la Figure 3.1,
- soit sous forme encyclopédique avec une définition de *Wikipédia*.

Sur la page de résultats et en fonction de la recherche, les termes connexes sont reliés entre eux par des liens avec des traits de couleurs différentes ; une palette de couleurs symbolise le rapport sémantique plus ou moins proche entre les termes. De nouvelles cartes peuvent être produites en quelques secondes lorsque l'utilisateur affine ou modifie sa recherche en cliquant sur un autre terme.

3.2 Traduction du besoin informationnel en langage de requêtes

Chaque jour presque 6 milliards de requêtes sont soumises au moteur de recherche *Google*[3] qui totalise environ 72% du trafic global[4] , *Baidu* à 15%, *Yahoo !Search* à 6%, *Bing* à 6% également et *AOL*, *Ask* et EXCITE à moins d'un % (données de Février 2014). Le langage de requêtes permet à l'utilisateur de formuler sa question sous forme de mots-clés dans un formulaire de recherche. Le formulaire de recherche consiste en libellés et

[2]http ://serelex.cental.be/
[3]http ://www.statisticbrain.com/google-searches
[4]http ://www.netmarketshare.com/search-engine-market-share.aspx

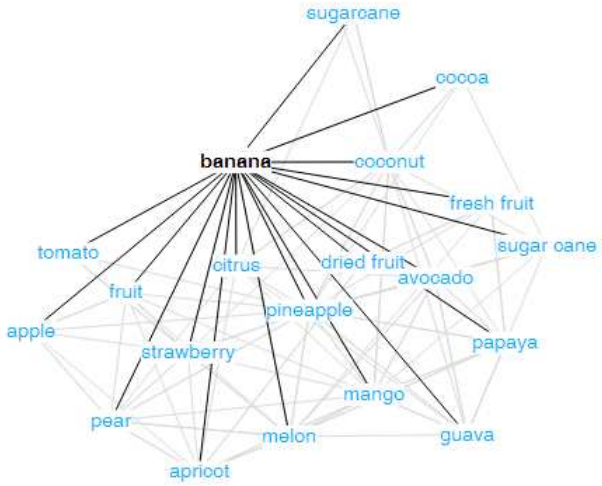

Figure 3.1 – Exemple de recherche par interface cartographique : Serelex

zones de saisie de valeurs de recherche. L'avantage de ce type de recherche est qu'il est accessible au plus grand nombre car à la fois simple d'accès par son interface simplifiée avec toutefois des possibilités de recherche avancée. Le langage de requêtes est efficace quand l'utilisateur a une bonne connaissance du vocabulaire de sa recherche. Il peut l'être moins lorsque l'utilisateur a moins d'expertise sur sa demande. Cette simplicité présente plusieurs inconvénients : *(1)* elle le rend peu adapté à des recherches sur des corpus généralistes et *(2)* elle le rend également moins appropriée à l'expression de requêtes plus longues car généralement peu encore traitées par les moteurs de recherches. Ces deux derniers points expliqueraient l'éventuelle perte de vitesse des moteurs de recherche qui ne proposent que ce type d'expression du besoin informationnel. La nécessité de passer à un système d'expression hybride parait indispensable.

3.2.1 Principes de l'interrogation en langage de requêtes

Dans le cas où l'utilisateur choisit de taper directement des termes dans un formulaire de recherche, également appelé « barre de recherche », il doit passer par une sélection et une saisie des éléments représentatifs de sa recherche. Cette formulation va dépendre à la fois du contenu recherché, des contraintes imposées par le moteur comme

les fonctionnalités ou la langue… mais aussi des circonstances de l'activité comme le temps disponible, le degré d'exigence de la tâche ainsi que du choix et de l'arrangement séquentiel des termes dans l'*équation de recherche*. Une équation de recherche correspond à la chaine de caractères entrée dans le formulaire de recherche, comme un ou des mots-clés ; rares sont les expressions entières en langage naturel [LI 10].

Pour un souci d'efficacité, la plupart des moteurs offre de nombreuses fonctionnalités pour affiner ces équations de recherche. Parmi les plus récurrentes :

- **les opérateurs de recherche** : ces opérateurs peuvent être booléens (ET, OU, SAUF), de proximité (NEAR) ou d'adjacence (ADJ).

- **les filtres de recherche** : par le biais de la recherche avancée, les utilisateurs peuvent ajouter des filtres particuliers comme la date de publication, la source d'origine, la langue, la zone géographique, le prix, le type des résultats ou la présence d'entité nommée comme une personne, un lieu, une organisation. Les filtres sont également appelés « facettes » qui désignent une façon particulière de présenter les données pour un utilisateur ou un groupe d'utilisateurs donné. Cette fonctionnalité permet aux utilisateurs de ne sélectionner qu'une partie de la collection de données en affinant l'affichage des résultats sur un ou plusieurs critères. On retrouve ce système sur de nombreux sites de commerce électronique par exemple. Certains moteurs permettent également de faire des requêtes plus restrictives : ne portant que sur un certain format de documents (filetype :pdf) ou sur un site ou domaine particulier (site :wikipedia). Ces restrictions peuvent être ajoutées aux mots-clés de la requête directement dans le formulaire de recherche.

Un algorithme est ensuite appliqué pour pour faire correspondre les mots-clés contenus dans la requête avec les documents en base documentaire. Ces algorithmes d'appariement font l'objet de très nombreuses investigations scientifiques. Les moteurs de recherche les plus simples se contentent de requêtes booléennes pour comparer les mots d'une requête avec ceux des documents. Les moteurs plus évoluées sont basés sur le paradigme du modèle vectoriel. : ils utilisent la formule *tf-idf* (*Term Frequency-Inverse Document Frequency)* qui permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Cette formule est utilisée pour construire des vecteurs de mots, comparés dans un espaces vectoriel, par une similarité cosinus.

Il existe d'autres nombreuses techniques pour améliorer les performances d'un moteur, notamment la méthode LSA qui permet d'établir des relations entre un ensemble de

documents et les termes qu'il contient, en construisant des concepts liés aux documents et aux termes. Ainsi, le terme « avion » sera automatiquement associé à ses mots proches tels que « transport aérien » ou « aéroport » ou à un nom de marque « Air France ».

Une stratégie de recherche idéalement et efficacement employée permettrait d'augmenter la pertinence des résultats et de minimiser le bruit documentaire. Cependant, la façon de transcrire la requête et l'étendue de leurs possibilités peuvent varier d'un système à un autre, compliquant ainsi la tâche de l'utilisateur. D'autres modules complémentaires aux requêtes permettent d'établir également un appariement amélioré. Notons parmi les modules les plus connus :

- **le correcteur orthographique ou la recherche floue** : lors de la recherche, certaines variations orthographiques peuvent être gérées par des lexiques de synonymie -notamment pour les noms propres- recensant les différentes orthographies possibles. Mais les fautes de frappe ou les erreurs orthographiques ne sont pas toutes prévisibles. Il convient alors d'effectuer ce que l'on appelle une *recherche floue* sur les termes en calculant la proximité des graphies de deux termes. Cette méthode permet ainsi de traiter les phénomènes d'insertion, de répétition, d'oubli ou de remplacement de lettres… Cette méthode peut être également utilisée notamment lorsque la requête n'a pas obtenu de réponses.

- **Le lemmatiseur** : cette opération consiste à opérer par réduction des mots en une entité première : le *lemme*. Le lemme regroupe toutes les variables du mot et ses dérivés (*i.e.* formes fléchies). Le traitement associé est la *lemmatisation*. Il s'agit de faire disparaître les différences morphologiques, *e.g.* les marques de nombre, de genre ou de conjugaison pour ne travailler que sur une *forme canonique*. Ainsi, un terme dans un formulaire de recherche tel que « cheval » pourra renvoyer comme résultats des documents traitement de « cheval » bien sur mais aussi de « chevaux », « chevalerie ». Ce traitement permet donc une correspondance plus importante entre les mots-clés et les termes issus des documents. Il peut, en contrepartie avoir des problèmes de bruit documentaire avec des résultats traitant de « chevalet ».

- **La liste de mots vides (*stop words*)** [RIJ 79] : le filtrage des mots vides consiste à éliminer les mots peu informatifs comme les prépositions, les articles, les pronoms (*e.g.* « de », « un », « les ») qui apparaissent très fréquemment et qui apportent uniquement de l'information sur le plan syntaxique et/ou logique. Certains problèmes

liés à des mots homographes (*e.g.* « son », « vers », « pas ») peuvent alors apparaître : ce sont des mots qui s'écrivent de la même manière et pour lesquels il faut alors différencier le mot vide de celui qui est utile de son contexte.

3.2.2 Les différents types de requêtes au sein des moteurs de recherche

Les besoins informationnels ne demandent pas tous le même traitement d'où la nécessité de les distinguer et de proposer des outils adaptés en conséquence. Ces outils peuvent être d'aide à la navigation, d'accès à du contenu multimédia ou encore de réalisation de transactions commerciales. Or, même si l'objectif des SRI consiste à faciliter l'accès à un ensemble de documents, on constate que la part laissée aux utilisateurs de ces systèmes est souvent limitée : son environnement ou sa tâche à effectuer peuvent être négligés. Pour cela, la tendance actuelle est de connaître au mieux l'utilisateur et notamment le but de sa recherche et / ou sa tâche à effectuer afin de mieux le servir. Cette connaissance permettrait de mieux prendre en compte l'hétérogénéité des utilisateurs ainsi que la variété des besoins. L'exploitation du contexte de la tâche de recherche fait l'objet d'une variété d'approches en RI contextuelle où le type de besoin inhérent à la requête peut être défini comme étant **informationnel** -lié à la recherche du contenu informationnel de documents-, **navigationnel** -lié à la recherche des sites d'accueil des personnes, organisations ou autres- ou **transactionnel** - lié à la recherche des services en ligne.

Nous présentons dans la partie suivante les principaux travaux des auteurs visant à mettre en place des modèles de requêtes guidés par la tâche de RI.

3.2.2.1 La typologie des requêtes de Broder

L'une des premières typologies des requêtes en fonction du but des utilisateurs a été proposée par [BRO 02]. Broder constate que le besoin informationnel des utilisateurs peut être de différentes natures :

- **les requêtes informationnelles** : le but de ces requêtes est tout simplement d'acquérir des informations supposées être présentes sur le Web. Seule la lecture des résultats est prévue. Les Les auteurs notent d'ailleurs que 15 % des recherches ont comme résultat désiré une collection de liens sur le sujet plutôt qu'un document.
- **les requêtes navigationnelles** : l'intention immédiate est de parvenir à un site particulier que l'utilisateur a en tête, soit parce qu'il l'a déjà visité dans le passé,

soit parce qu'il suppose qu'un tel site existe. Ce type de recherche est également appelé *recherche de l' élément connu* et n'a en principe qu'un seul *« bon »* résultat. Ainsi une recherche comme *« El Djazair News »* (nom d'un journal algérien) aura probablement comme résultats cibles l'un des sites suivants : le site en arabe à l'adresse suivante : www.djazairnews.info/ ou le site en français : www.presse-dz.com/presse-algerie/djazair-news/.

- **les requêtes transactionnelles** : l'intention de ces requêtes est d'atteindre un ou plusieurs sites permettant d'effectuer des transactions *e.g.* faire des achats, télécharger des fichiers, des images, des musiques, accéder à certaines bases de données ou trouver des serveurs pour jouer en ligne. Par exemple, un utilisateur soumettant la requête « promo PC portable 15 pouces » désire certainement accéder à un site de comparateur de prix d'ordinateurs portables de taille 15 pouces.

Afin d'estimer la proportion de chaque type de requêtes, Broder a sélectionné une palette de 1000 requêtes à partir du moteur de recherche AltaVista. L'auteur a toutefois supprimé de son échantillon de requêtes, celles qui n'étaient pas formulées en anglais ainsi que celles qui avaient un caractère sexuel. Sur ces 1 000 requêtes, 400 ont été retenues (ils ont exclu les requêtes non formulées en anglais ainsi que celle qui présentent un caractère sexuel) ; il en ressort que près de la moitié des requêtes (48 %) sont informationnelles, 30 % sont transactionnelles et 20% sont navigationnelles. Notons que les requêtes qui n'étaient ni transactionnelles ni navigationnelles ont été considérées comme ayant une intention informationnelle, or cette conclusion pourrait être discutée car les résultats pourraient en être faussés.

3.2.2.2 La typologie des requêtes de Rose et Levinson

[ROS 04] ont proposé une typologie des requêtes dont l'objectif est de proposer un cadre conceptuel permettant d'identifier et d'organiser un ensemble des types canoniques de buts des utilisateurs. Les auteurs ont organisé et hiérarchisé les requêtes à partir de la trichotomie de Broder mais avec une altération :

- **les requêtes ayant un but navigationnel** : volonté de la part de l'utilisateur d'être redirigé vers la page d'accueil d'une institution ou d'un organisme. Pour être considérée navigationnelle, la requête doit avoir un site web référent que l'utilisateur avait déjà en tête. Pour cette raison, la plupart des requêtes consistent en

des noms de sociétés, d'universités, d'organisations bien connues par l'utilisateur (« Université de la Sorbonne ».

- **les requêtes ayant un but informationnel** : concentrées sur l'objectif de l'utilisateur de se procurer des informations sur le sujet de la requête. Elles sont de plusieurs types : *(a)* celles qui sont **directes**, quand l'utilisateur cherche une connaissance particulière sur un sujet donné ; *(b)* celles qui sont **indirectes**, quand l'utilisateur veut apprendre sur un sujet (*« tell me about X »*) ; *(c)* celles qui sont de type **conseil** : quand l'utilisateur veut obtenir des conseils, des avis, des suggestions ou des instructions sur un sujet particulier comme sur des forums de discussions ; *(d)* celles qui concernent la **localisation d'une information en particulier**, quand l'utilisateur cherche à accéder à un service ou un produit particulier ; *(e)* celles qui concernent la **constitution de listes**, quand l'utilisateur veut obtenir une liste de sites web qui potentiellement pourraient l'aider à répondre à sa question (*e.g.* une liste de sites qui donnent les informations sur les activités à pratiquer dans une région).

- **les requêtes ayant un but de ressources** : représentent des requêtes ayant l'objectif d'obtenir quelque chose autre que l'information disponible sur le site web (*e.g.* paroles de chansons, recettes, patrons de couture, etc.). Elles concernent : *(a)* **le téléchargement** : quand l'utilisateur télécharge un outil utile pour l'utilisation d'une autre ressource (*e.g.* logiciels *Peer-to-Peer* comme *Kazaa Lite*) ; *(b)* **le loisir** : quand l'utilisateur a pour objectif de se divertir tout simplement par des éléments d'affichage disponible sur la page de résultat (*e.g.* vidéo, clips) ; *(c)* **l'interaction** : quand l'utilisateur interagit avec une ressource à l'aide d'un autre programme / service disponible sur le site web qu'il a trouvé (*e.g.* outil de traduction) ; *(d)* **l'obtention** : quand l'utilisateur obtient une ressource à imprimer ou à consulter. L'objectif ici n'est pas d'apprendre quelque chose mais de pouvoir utiliser la ressource elle-même.

Les requêtes ayant un but de ressources renvoient d'une manière plus générale aux requêtes transactionnelles de Broder, l'utilisateur souhaitant accéder à une ressource présente sur une page Web, plutôt que de consulter directement une information.

Les auteurs ont testé leur classification sur un corpus d'environ 500 requêtes en anglais du moteur de recherche *Altavista*. Conscients de la difficulté d'inférer le but sous-jacent aux requêtes simplement à partir du texte des requêtes elles-mêmes, les auteurs

s'appuient ainsi sur quatre types d'information accessibles : le texte de la requête, les résultats retournés par le moteur de recherche, les résultats cliqués par l'utilisateur et les interactions ultérieures de l'utilisateur. Environ 62 % des requêtes ont un but informationnel, 25 % ont un but de recherche de ressource et 13 % ont un but navigationnel. Tout comme dans la typologie de Broder, les requêtes informationnelles dominent le reste des requêtes.

3.2.2.3 La typologie des requêtes de Jansen, Booth et Spink

[JAN 08] proposent une méthodologie pour classer les intentions de recherche des utilisateurs sur le Web en s'inspirant de la trichotomie des requêtes (*i.e.* navigationnelles, informationnelles et transactionnelles) pour fonder leurs recherches. Ils définissent pour cela un ensemble de traits caractéristiques pour chaque type de requêtes qu'ils implémentent dans une tâche de classification automatique. Ils utilisent 4 250 656 requêtes du méta-moteur de recherche *Dogpile* (ensemble des requêtes des moteurs de recherche*Excite, AlltheWeb* et *AltaVista*) entre 1997 et 2002. Leurs résultats montrent que 80 % des requêtes sur le Web sont de nature informationnelle, e les 20 % restant se répartissent équitablement en requêtes navigationnelles et transactionnelles. A noter que dans cette étude les résultats sur le nombre de requêtes informationnelles (80 %) sont beaucoup plus importants par rapport aux autres études précédemment citées -40 % pour Broder et 62 % pour Rose et Levison-. Ceci peut être expliqué par la taille du corpus, ce qui relativiserait les résultats obtenus sur les requêtes navigationnelles et transactionnelles des deux études de Broder et Tose et Levinson.

La distinction de ces différents types de requêtes a donc permis de cerner le domaine d'accès contextuel à l'information guidé par la tâche de recherche : informationnelle, navigationnelle ou transactionnelle. L'objectif sous-jacent de la détection du type de besoin informationnel à travers l'expression de la requête est de mettre en place un modèle de RI orienté tâche pour ensuite pouvoir orienter la sélection du type d'informations à fournir en résultats.

La thèse de [DAO 09] porte notamment sur la prédiction du type de besoin des requêtes, traduisant la nature de la tâche de recherche, en exploitant les caractéristiques morphologiques (des requêtes) ainsi que le profil et le contexte de la session. Son travail aborde plus particulièrement les requêtes ambigües, dites multi-facettes, qui présentent des caractéristiques associées à plusieurs types de besoins. En effet, la plupart des approches exploitent seulement des caractéristiques morphologiques de la requête

[KAN 03], [KAN 05] ou des indicateurs de comportement de l'utilisateur (clics, données de consulation de la page, nombre de documents consultés…).

Le postulat de départ du travail de [DAO 09] est que l'utilisation isolée des caractéristiques de la requête est insuffisante pour prédire son type. Cela est particulièrement vrai dans les cas de multi-facettes ou de requêtes ambigües. Par exemple, d'après les typologies que nous venons de voir, l'existence d'un verbe dans les termes de la requête est une caractéristique de requêtes qui peuvent être à la fois informationnelles et transactionnelles (*e.g* « voyager » pourrait appartenir aux deux catégories). Sa méthode permet notamment de renforcer la prévision de la classification d'une requête grâce au type de profil de la session. Des procédés de reformulations et d'enrichissements de requêtes incluant des tâches donnent un gain de performance significatif.

3.2.3 Degrés différenciés d'explicitation et de difficulté des requêtes

Parallèlement aux typologies des requêtes en fonction du but des utilisateurs, [STR 08] font l'observation que les requêtes varient considérablement du point de vue de leur degré d'explicitation. Ils distinguent deux types de requêtes selon que le but soit implicitement ou explicitement exprimé.

3.2.3.1 Des requêtes plus ou moins explicites

[STR 08] distinguent :

- les requêtes dont le but est explicitement exprimé : ce sont des requêtes qui décrivent avec précision leur intention de recherche, *i.e.* pouvant être reliées à un but spécifique, de manière reconnaissable et non ambiguë. Exemple : *« acheter une voiture »*, *« réparer une voiture »*, *« aller à Miami »*,

- les requêtes dont le but est implicitement exprimé : ce sont des requêtes où il est difficile d'obtenir le but spécifique des intentions de recherche car elles sont plus floues voire mal formulées par l'utilisateur. Exemple : *« voiture »* ou encore *« voyage »*. Ces requêtes doivent alors être affinées afin d'obtenir des résultats pertinents.

Afin de mieux comprendre les différences entre ces deux niveaux d'expression, les auteurs extraient automatiquement un ensemble de requêtes intentionnelles explicites

et implicites d'un journal de requêtes (appelés plus communément *logs* de requête) du moteur de recherche *AOL*. [5]

Sur un ensemble de 279 260 requêtes extraites, il apparaît ainsi qu'environ la moitié sont de type intentionnel explicite. Par exemple, les alertes Google permettent à l'utilisateur de saisir une requête « profil » servant de critère de sélection de l'information (web, actualités, etc.) L'utilisateur est alors averti par email lorsque de nouveaux résultats correspondant aux termes qu'il a renseigné sont publiés (Web, actualités, etc.). Ces techniques sont intéressantes pour aider à la construction d'un profil « contrôlé » d'un utilisateur mais imposent un effort supplémentaire à l'utilisateur : celui de spécifier explicitement ces centres d'intérêts.

Pour les deux types de requêtes (implicites et explicites), il est important que le système de recherche d'informations utilisé puisse lever l'ambiguïté ou approfondir certaines recherches. Ex : un *« magasin de voiture »* peut être précisé en *« magasin de réparation de voiture » « trouver un magasin d'occasions de voiture », « magasin d'achats de voiture »* qui correspondent à des besoins différents.

3.2.3.2 Des requêtes plus ou moins difficiles

[MOT 11] propose également une approche qui préconise des mécanismes adaptés du SRI et des traitements spécifiques en fonction du type de la requête. Elle regroupe les besoins d'informations selon que les requêtes sont identifiées comme *faciles* ou *difficiles* selon le système. Évaluer la difficulté d'une recherche présente au moins deux avantages : une interaction homme-machine et une évaluation du système. Si le système n'est pas en mesure de déterminer la difficulté d'une requête, il peut donc demander à l'utilisateur de redéfinir sa demande soit en la reformulant soit en donnant plus d'informations permettant de désambiguïser les termes de la demande. Si un terme n'apparaît pas dans le corpus, le système peut également demander à l'utilisateur de remplacer le ou les terme(s) employé(s) ou d'utiliser des synonymes.

Parmi les indicateurs de la difficulté des requêtes, notons :

- la fréquence des termes de la requête dans l'ensemble des documents et le nombre de sens dans ces termes, d'après [DEL 00],

[5] Le *logging* ou l'historique des événements désigne l'enregistrement séquentiel dans un fichier ou une base de données tous les événements affectant un processus particulier (application, activité d'un réseau informatique,etc). Le journal (en anglais *log file* ou plus simplement *log*), désigne alors le fichier contenant ces enregistrements.

- l'entropie relative entre le modèle de la langue de la requête et le modèle de la collection, appelée le score de clarté *clarity score*, par [CRO 02]). Ce score de clarté mesure en effet l'ambiguïté d'une demande par rapport à la collection de documents et montre qu'il est corrélé positivement avec la précision moyenne des documents du programme d'évaluation TREC. Cela permet également d'identifier les demandes inefficaces sans détériorer pour autant les scores d'appariement d'un système.

- le nombre d'entités nommées pour [MAN 02],

- les caractéristiques linguistiques extraites des requêtes par [MOT 05]. Ces caractéristiques sont au nombre de seize et peuvent être extraites de façon automatique pour être étudiées. Elles correspondent à trois niveaux linguistiques : morphologique, syntaxique et sémantique. Parmi les traits morphologiques, notons : le nombre de mots [NBWORDS], la longueur moyenne du mot [LENGTH], le nombre moyen de morphèmes par mot [MORPH], le nombre moyen de suffixes [SUFFIX], le nombre moyen de noms propres [PN], la moyenne des acronymes [ACRO], le nombre moyen des valeurs numériques comme les dates ou les quantités [NUM], le nombre moyen des *tokens* inconnus [UNKNOWN]. Les traits syntaxiques sont mesurés par la moyenne des conjonctions [CONJ], la moyenne des prépositions [PREP], la moyenne des pronoms personnels [PP] et deux autres mesures de complexité linguistique [SYNTDEPTH] et [SYNTDIST]. Enfin, le trait sémantique est caractérisé par la moyenne des valeurs polysémiques [SYNSETS].

- le score de cohérence [HE 08], qui s'appuie sur la cohérence d'une ensemble de documents ; un ensemble de documents est cohérent si une proportion de documents se ressemblent. L'ensemble des documents retrouvés par un terme de la requête correspond alors à la cohérence associée à ce terme.

- le score de couverture d'un besoin (CTS[6]) [LAN 08] qui estime si le besoin exprimé au travers d'une requête est couvert par l'ensemble des documents retrouvés. Le CTS est haut si les termes de la requête sont fréquents.

[6]Covering Topic Score

3.2.4 Les reformulations et les expansions de requêtes

Depuis quelques années des travaux sont menés pour produire des reformulations ou des expansions de requêtes [TAN 05], [GAR 10], [VEN 09] notamment grâce à des lexiques, des thésaurus ou encore des ontologies de domaines [AIM 10]. Une expansion de requêtes peut être vue comme un traitement pour *élargir* le champ de recherche pour cette requête. Une requête étendue va contenir plus de termes reliés : soit des termes synonymes de même niveau hiérarchique que celui saisi dans la requête, soit de niveau hiérarchique différent en faisant des inférences et en utilisant des connaissances stockées dans un thésaurus ou dans une ontologie par exemple. En utilisant le modèle vectoriel, par exemple, et en sélectionnant des documents selon un seuil de similarité entre un document et une requête, plus de documents pertinents seront repérés (engendrant aussi plus de bruit). Cette méthode devait avoir comme résultat une augmentation du taux de rappel. Pour le modèle booléen par exemple, une expansion de requête se définit par une forte relation entre un terme ou un syntagme A et un terme ou un syntagme B ou une relation d'implication $B \Rightarrow A$. Si A apparaît dans une requête booléenne, alors on remplace A par $(A \cup B)$ et l'expansion se fait par des synonymes. Si B est un bon substitut de A, alors la requête étendue ne change pas la sémantique de la requête initiale et en général comme la requête booléenne n'est pas pondérée, il n'y aura pas non plus de pondération sur le terme ajouté[7]. Cette méthode de pondération a été expérimentée par plusieurs chercheurs dont [VOO 94]. [XU 96] et al. montrent que l'expansion de requêtes est adaptée aux requêtes courtes alors que la relaxation de requêtes (suppression de termes) est efficace pour les requêtes longues.

Lorsque les utilisateurs étendent ou relaxent leurs requêtes eux-mêmes, dans les cas de reformulation, ils ajoutent ou suppriment en moyenne un à deux termes et environ 35% des requêtes modifiées sont sans changement du nombre total de mots [JAN 00a].

Notons également que certains auteurs s'intéressent à l'historique des requêtes pour améliorer les performances des moteurs de recherche. En cas de reformulation, le besoin informationnel pourra être enrichi et clarifié en prenant en compte les termes précédemment saisis [SHE 03] ; le postulat sous-jacent est que les requêtes précédentes fournissent également un contexte intéressant à prendre en compte. [LEV 13a] distingue d'ailleurs différentes sessions de requêtes pour un même utilisateur lorsque les environnements sémantiques sont trop éloignés. Si ces environnements sémantiques sont proches, alors

[7] Dans le modèle vectoriel au contraire, le terme ajouté B reçoit le même poids que le mot initial A en relation à un facteur. Ce facteur peut être fixe ou bien déterminé selon le nombre de B en relation avec A

l'auteur les utilise comme contexte de requêtes -il utilise notamment les similarité des résultats pour détecter automatiquement des sessions de recherche [LEV 13b]. Dans [SHE 03], les documents retrouvés par plusieurs formulations du besoin sont alors considérés comme potentiellement plus pertinents. Ces travaux sont complétés par ceux de [SHE 05]. Les auteurs combinent l'historique des requêtes avec les clics effectués sur les résultats (*snippet*[8]) comme retour implicite des utilisateurs. Ces auteurs montrent que la combinaison de ces deux critères améliore en moyenne la précision ; *a contrario* l'utilisation des historiques seule est peu probante. Dans la même continuité, les travaux de [TAN 06] reprennent les deux facteurs précédemment étudiés dans les historiques (requêtes et documents retrouvés) auxquels ils rajoutent les liens sélectionnés par l'utilisateur. C'est une stratégie à long terme qui permet d'obtenir des informations riches sur le profil et le besoin informationnel des utilisateurs. Les auteurs montrent notamment que prendre l'historique à long terme améliore également les performances d'appariement du système. D'autres travaux reposent sur le regroupement des termes des requêtes par catégories thématiques plus générales [PU 02] ou selon une taxonomie [CHU 03].

3.3 Traduction du besoin informationnel en langage naturel

On regroupe dans le concept de Traitement Automatique du Langage Naturel (TALN) l'ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication [YVO 07].

3.3.1 Principes de l'interrogation en LN

Les moteurs de recherche qui proposent le langage naturel (LN) comme moyen d'interrogation voire d'interaction dans certains cas de systèmes de Q|R, doivent être théoriquement capables d'interpréter des questions posées avec des mots de la langue courante. L'objectif sous-jacent de ces moteurs est de proposer un mode d'interrogation non contraignant pour l'utilisateur, afin qu'il puisse exprimer librement son besoin informationnel. L'utilisation de la LN devrait faire face aux problèmes de complexité d'utilisation des SRI ; au contraire des langages artificiels ou formels, la LN se caractérise par son absence de conception spécifique et consciente. Ce mode d'expression devrait en théorie accroître la précision en apprivoisant l'ambiguïté *via* une compréhension enrichie

[8] Le snippet est la courte présentation qui présente une page dans les résultats de recherche. Il comprend souvent un titre, une description et parfois des informations supplémentaires (images, votes, prix ...).

du contexte ; le besoin informationnel est en effet davantage contextualisé (la formulation plus proche et plus longue) que dans une équation de recherche. Aujourd'hui, de nombreuses applications industrielles comme la traduction automatique, les interfaces en langage naturel ou les systèmes de Q|R sont utilisées.

3.3.1.1 Les applications du TALN

Parmi les nombreuses applications du TALN, nombreux outils ont vu le jour : correcteur orthographique, traduction automatique, résolution d'anaphores, annotation sémantique, reconnaissance d'entités nommées, détection de co-références ou encore reformulation et paraphrasage en ce qui concerne les textes et les énoncés. Les applications de l'utilisation de la LN peuvent être multiples et représentent un fort potentiel et attrait industriel notamment pour les services proposés aux clients. Les applications comme la reconnaissance automatique de la parole, la synthèse de la parole, la reconnaissance vocale, le traitement de la parole en sont des exemples. Ainsi, il n'est pas rare de formuler une demande de type « trouver le meilleur restaurant italien près de Seattle Washington » plutôt que « restaurant italien Seattle WA ». De telles demandes ont été multipliées et développées notamment avec l'augmentation des usages des applications de reconnaissance vocale des smartphones [XU 13].

Ces applications sont très présentes dans le cadre de Services Après-Vente de France Télécom ou EDF (voir notamment le récent article [SUI 13] sur le sujet). Nous pouvons citer à titre d'exemple le serveur vocal d'*Orange Labs* lancé en avril 2011 : cet outil permet au client de l'opérateur téléphonique passant par le 39-00, le 10-13 ou le 3000, d'exprimer librement sa demande. Un dictionnaire réactualisé de 3 000 formules permet de couvrir le champ des demandes. Le taux de satisfaction garanti serait, selon *France Télécom*, de plus de 80 %. *France Télécom* vend également cette technologie au secteur privé. La Banque postale et le Crédit agricole l'ont notamment adoptée. On retrouve ces mêmes technologies sur des sites commerciaux comme voyage-sncf.com ou celui d'*ikea* par exemple. Fort de ces expériences très positives, la RATP, la SNCF, le LIMSI-CNRS et la société *vecsys* ont noué un partenariat pour mener à bien le projet *SATIM*[9]. Ce projet a pour objectif la réalisation du prototype d'un Serveur Vocal Interactif (SVI) en langage naturel, qui donne accès par téléphone, à la recherche d'itinéraires multimodaux (réseau ferré RATP et SNCF, tram et bus) pour tout trajet immédiat ou différé en Ile de France. Nous nous axerons dans le présent travail exclusivement sur l'application du TALN dans

[9]Serveur Vocal d'Accès à la recherche d'iTInéraires Multimodaux en Ile de France

les moteurs de recherche.

3.3.1.2 Les moteurs de recherche proposant l'interrogation en LN

Nous nous intéressons ici plus spécifiquement à l'usage de la LN dans les moteurs de recherche. D'un point de vue indexation, il s'agit de ne plus aborder la demande comme une suite de mots mais comme un ensemble fortement structuré qui permet de communiquer des informations grâce à l'utilisation de techniques de TAL. Les mots de la demande sont organisés comme des entités linguistiques à part entière *i.e.* comme des unités susceptibles de posséder plusieurs sens, de subir des variations de formes, de structures et d'entretenir des relations avec d'autres entités. Cette richesse et cette complexité doivent être prises en compte pour une meilleure compréhension de la demande. [MOT 13] rappelle que les deux disciplines -TAL et RI- sont imbriquées et qu'elles doivent avancer par projets multi-disciplinaires pour bénéficier de leurs apports mutuels nombreux. Pour [XU 13] pour qui les demandes informationnelles s'établissent de plus en plus en LN et à l'instar de Google qui lance en septembre 2013 son nouvel algorithme baptisé *« Hummingbird »*, le langage d'interrogation en LN a un regain d'intérêt ces derniers temps. Plusieurs autres moteurs de recherche se sont lancés -avec plus ou moins de réussite- dans l'interrogation en langue naturelle. Nous avons testé quelques moteurs proposant de mode d'interrogation avec la question suivante : « Who is the president of the Senate in France ? » ou « Qui est le président du Sénat en France ? » selon la langue autorisée. Ce test a été effectué en septembre 2013 ; les résultats sont présentés dans la Figure 3.2, page 55.

C'est liste n'est pas exhaustive, d'autres moteurs intègrent les techniques de TAL dans des outils de calcul comme *Wolfram Alpha*, d'autres dans des systèmes de paraphrase *Inbenta*. En effet, *Wolfram Alpha* (www.wolframalpha.com) est un outil de calcul en langage naturel développé par la société internationale Wolfram Research. Il s'agit d'un service internet qui répond directement à la saisie de questions factuelles en anglais par le calcul de la réponse à partir d'une base de données. A la question « Who is the president of the Senate in France ? », *Wolfram Alpha* identifie bien comme *Input representation* : President of the Senate (qui correspond à un libellé *leadership position)* et comme résultats des noms, prénoms de présidents du Sénat, la durée de leurs mandats, ceci pour deux pays : la Pologne et la Grèce mais la France n'apparaît pas dans les résultats. Il semble que le moteur ait bien interprété la question mais que cette information ne soit pas enregistrée dans leur base de données. Il conviendrait alors de le mentionner

à l'utilisateur.

Inbenta (http ://www.inbenta.com/) est un moteur de recherche spécialisé dans les secteurs de la Banque et des Assurances en français et en espagnol. Son système permet d'interpréter les Q|R en LN et notamment de générer des paraphrases. Son système est basé sur des fonctions lexicales permettant de désigner formellement les relations entre les mots et de les formaliser (relations synonymiques, verbales, etc.). Le module d'interrogation n'est pas en test en ligne mais une démonstration lors de TALN'2013 [QUI 13] nous a tout de même permis de constater que ce module permettait de générer des paraphrases à partir de questions des utilisateurs. Sa notoriété n'est pas encore très importante mais sa spécificité (secteur bancaire et assurance) peut être un atout important pour ce genre de traitement.

L'appellation peut également être différenciée selon les auteurs : on rencontre dans la littérature le nom de *moteurs de recherche sémantique* dans le sens où l'utilisateur fournit au moteur une phrase/une question destinée à désigner l'objet sur lequel il tente de recueillir de l'information. Sur le fond, ces moteurs convergent vers une dimension sémantique de la demande ; les mots-clés figurent dans un ensemble formé et articulé qui prennent sens en fonction de leur position et de leur contexte. Pour cela, plusieurs analyses linguistiques sont nécessaires.

3.3.1.3 Les segmentation et les classifications des demandes en LN

Des travaux similaires ont vu le jour sous le nom de *Question Answering* (QA) dont l'objectif est de fournir des réponses concises (et non la totalité du document) à des questions en LN du type « Combien de filiales a Coca-Cola dans le monde ? ». Ce sont en effet des questions factuelles qui appellent en réponse des éléments de réponses précis et courts. L'objectif est donc d'apporter des réponses directes à la demande plutôt que la liste de documents à consulter. En grande partie, le succès d'une telle approche repose sur une compréhension solide de l'intention de la demande ; il faut donc comprendre la question factuelle en langage libre afin de déterminer les contraintes qu'impose une telle question et envisager les réponses possibles. Or, il y a relativement peu de travaux sur la compréhension sémantique des requêtes web. [MAN 09] and [LI 10] ont travaillé sur la structure sémantique des requêtes. [LI 10] en particulier a travaillé à partir des syntagmes nominaux sur l'intention de tête (attribut) couplés à un certain nombre de modificateurs (valeurs d'attributs). Ainsi, la requête « Distribution Alice au pays des merveilles 2010 » est constitué d'une tête d'intention « Distribution » et de deux modificateurs d'inten-

tions :«Alice au pays des merveilles » et « 2010 ».

Certains travaux, basés sur les requêtes cherchent à attribuer des étiquettes (*querry tagging*) sémantiques pour chaque terme de la requête une catégorie pré-définie [LI 09], [MAN 09] [SAR 10]. Ainsi dans la requête « Canon digital camera silver », le terme « Canon » sera étiqueté comme une marque ; « digital camera » comme le modèle et « silver » comme l'attribut. [PAS 07] donne des attributs comme « ville capitale », « President » pour la classe Pays ou « coût», « manufacturing », « effets secondaires » pour la classe « Drogue ». Une application immédiate de leur recherche est de comprendre les requêtes qui demandent des informations factuelles de certains concepts comme les attributs « coût » ou « effets secondaires » d'un nom de drogue comme « aspirine » et donc de fournir des réponses plus appropriées. Pour [XU 13], pour comprendre l'intention sémantique de la demande, le modèle ne doit pas être capable de lister tous ses constituants (type d'entité, lieu, date, etc) mais seulement de comprendre la structure de chacun. Ainsi, dans l'exemple « Combien d'états la rivière Colorado parcourt-elle ? », la sortie de l'analyse sémantique doit être : « comptabilise (état(parcourir(rivière(const(Colorado)))) ». L'objectif est d'étudier la compréhension de l'intention de la demande et d'en reconstruire une représentation hiérarchique. L'avantage d'un tel traitement des demandes est double : premièrement en reconnaissant certaines classes d'éléments comme les entités numériques ou les noms propres, le système peut étiqueter différemment ces entités et donc avoir des traitements différenciées et donc de meilleure qualité ; deuxièmement l'analyse des demandes permet de mieux cibler les potentielles réponses à donner. Les difficultés de cette tâche sont *(a)* tout d'abord la complexité des demandes très courtes et *(b)* la correspondance entre les termes utilisés dans les demandes et ceux utilisés dans les documents [CAR 12]. Une des étapes les plus importantes dans ce processus est d'analyser la question qui permet de déterminer le « type » de réponse recherché. Pour cela, plusieurs auteurs ont proposé des classifications comme [LI 02] à partir des questions de TREC-10 qui propose une taxonomie sémantique des réponses à fournir. La taxonomie comprend six grandes classes : abréviation, entité, description, Personnalité, localisation géographique et valeur numérique et cinquante classes plus fines.

3.3.2 Analyses linguistiques inhérentes à la compréhension de la langue

Une modélisation sous forme de simplifications structurantes est nécessaire pour rendre la LN interprétable par une machine. L'analyse des différents énoncés -nous nous

situons ici plus particulièrement sur les énoncés écrits- entretient avec la linguistique des rapports complexes. Il est à la fois nécessaire d'étudier des propriétés langagières des énoncés écrits mais aussi et surtout les conditions dans lesquelles ils s'opèrent. Du niveau morphologique, aux niveaux syntaxique et sémantique, plusieurs niveaux d'analyses linguistiques sont utilisés pour améliorer les mécanismes traditionnels de représentation des contenus des documents et requêtes. Ces différentes analyses permettent notamment de palier les problèmes inhérents de compréhension d'une langue : les problèmes d'ambiguïté du langage mais aussi de contextualisation des énoncés [YVO 07].

3.3.2.1 Le niveau morphologique

Le niveau morphologique est l'étude de la formation des mots et des variations de forme. Il concerne la reconnaissance d'unités linguistiques de base : les morphèmes *i.e.* les plus petites unités de sens isolées dans un énoncé. Les morphèmes peuvent être soit lexicaux soit grammaticaux. Les morphèmes lexicaux -ou *lexèmes*- appartiennent à une classe ouverte *i.e.* on ne peut pas dénombrer entièrement les lexèmes et les racines d'une langue alors que les morphèmes appartiennent à une classe finie *i.e* dont on peut compter le nombre de suffixes et de préfixes.

L'analyse morphologique peut être vue sous trois angles :

- **la morphologie flexionnelle** étudie la modification de la forme de référence d'un mot -nommée *forme canonique* ou *lemme*-,*i.e* la forme d'un mot sans ses marques dites de *flexions*. Il s'agit des flexions de genre, de nombre pour les noms, les adjectifs et les pronoms et des flexions de personne, de nombre, de temps, de mode et de voix pour les verbes.

  Exemple pour « étude » :

  « études », « étudié », « étudiée », « étudiés », « étudiées », « étudier », « étudie », « études », « étudions »… « étudiais », « étudiait »…

- **la morphologie dérivationnelle** étudie les dérivations à partir d'un radical donné ; les préfixes qui sont antéposés au radical comme « dé », « re », « a » et les suffixes qui sont postposés au radical comme « ment », « able », « age », « al », « eur ». Exemple : « nation » $\rightarrow$ « national » ou encore « lune » $\rightarrow$ « alunir ».

- **la morphologie compositionnelle** étudie de la création de nouveaux mots à partir de plusieurs radicaux. Exemples : Nom + Nom « homme-grenouille », Nom +

Adjectif « feu rouge », Verbe + Nom « porte-bagages », Nom « de » Nom « pomme de terre ».

Cette analyse a pour objectif de procéder à un recodage des diverses variantes à une forme unique. Si on s'oriente vers la morphologie flexionnelle, cette forme unique sera le *lemme* -forme unique débarrassée de ses flexions-. Le traitement associé est la *lemmatisation*. Les morphologies dérivationnelle ou compositionnelle auront pour principal résultat la normalisation des formes autour d'une *racine*[10] ou d'un *radical*[11].

Il y a également d'autres approches moins sophistiquées qui sont considérées comme des « approximations » de traitements linguistiques comme le *stemming i.e.* racinisation [GAU 97]. Cette approche cherche à rassembler les différentes variantes d'un mot autour de son *stem* (*i.e.* une pseudo-racine). Cette procédure traite à la fois des cas relevant de la flexion et de la dérivation. Les techniques utilisées pour procéder à la racinisation reposent généralement sur une liste d'affixes[12] et sur un ensemble de règles de désuffixation construites *a priori*. Elles permettent de retrouver le *stem* d'un mot. L'avantage de ces outils (*stemmer*) réside dans leur simplicité : ces outils gèrent en même temps les morphologies dérivationnelle et flexionnelle. Le stemming est d'ailleurs la méthode la plus utilisée dans les moteurs de recherche [LOU 04]. *A fortiori*, on peut regretter l'absence de contraintes linguistiques fortes qui engendre des erreurs de :

- sur-racinisation comme dans l'exemple suivant : le *stem* « nat » regroupe à la fois « nature » et « nation »
- ou de sous-racinisation : le *stem* « adapt » qui empêche le regroupement des formes « adapter » et « adaptation ».

Même si cette approche est très commune, l'impact du *stemming* sur les performances des SRI est en général mitigé.

[10] On appelle *racine* l'élément de base, irréductible, commun à tous les représentants d'une même famille de mots à l'intérieur d'une langue ou d'une famille de langue. La racine est obtenue par élimination de tous les affixes et désinences, elle est porteuse des sèmes essentiels, communs à tous les termes constitués avec cette racine. La racine est donc une forme abstraite qui connaît des réalisations diverses. [DUB 01] : p.395

[11] On appelle *radical* une des formes prises par la racine dans les réalisations diverses des phrases. La racine « ven » a deux radicaux « ven » et « vien » qui se réalisent dans les formes « venons » , « venue », « vienne », « viennent ». Une racine peut n'avoir qu'un seul radical, en ce cas racine et radical se confondent [DUB 01].

[12] Un affixe est un morphème en théorie lié qui s'adjoint au radical ou au lexème d'un mot.

La tâche se complexifie avec notamment les noms composés, très fréquents dans les documents techniques (exemples : « système expert », « réseau de neurones », « système distribué », « langage objet » ou encore « base de données »). C'est une analyse morpho-lexicale qui permet d'identifier ces mots composés et autres expressions idiomatiques dans des lexiques. Il faut ainsi figer leurs formes dans des lexiques notamment électroniques.

Parmi les lexiques électroniques, citons :

1. le Dictionnaire électronique des formes fléchies du Français (DELAF), développé par l'Institut d'Electronique et d'Informatique Gaspard-Monge,

2. Le Lexique Électronique des Formes Fléchies du Français (LEFFF), développé par l'INRIA et le Laboratoire Bordelais de Recherche en Informatique (LABRI) depuis 2003.

Pour traiter les problèmes de sur- ou sous-racinisation ainsi que pour traiter la complexité des noms composés, les moteurs de recherche peuvent :

- soit faire appel à des dictionnaires (ou lexiques) électroniques qui sont composés d'une liste de termes ou expressions avec leurs synonymes et leur appartenance grammaticale ; c'est une tâche gourmande en terme de temps de traitement,

- soit mettre en place des traitements statistiques à partir d'un corpus de textes pour la reconnaissance de certains syntagmes, sans avoir recours aux dictionnaires.

Les lexiques électroniques et les analyseurs morphologiques constituent donc un maillon important de la chaîne de traitement.

Un des traitements morphologiques est la tâche d'étiquetage morpho-syntaxique (*Part-of-Speech tagging* ou *POS tagging*) : elle consiste à identifier pour chaque mot sa classe morpho-syntaxique à partir de son contexte et à partir de connaissances lexicales. L'étiquetage morpho-syntaxique peut être vu comme la composition de 3 fonctions : (1) la segmentation du flux de caractères en mots, (2) l'étiquetage *a priori* (hors-contexte) des mots au moyen des informations lexicales qui associe toutes les étiquettes possibles pour un mot donné, (3) la sélection en fonction du contexte du mot, de l'étiquette la plus pertinente parmi celles identifiées par l'étiquetage *a priori*.

Parmi les outils de traitement morphologiques, nous pouvons citer :

- **TreeTagger** : *TreeTagger* est un outil qui permet d'annoter un texte avec des informations sur les parties du discours (genre de mots : noms, verbes, infinitifs et

particules) et des informations de lemmatisation. Il a été développé par l'Université de Stuttgart. Les ressources linguistiques sont disponibles pour de multiples langues dont le français. *TreeTagger* permet l'étiquetage de l'allemand, l'anglais, le français, l'italien, l'allemand, l'espagnol, le bulgare, le russe, le grec, le portugais, le chinois et du français ancien. Il est adaptable à d'autres langues si des lexiques et des corpus étiquetés manuellement sont disponibles. La présence d'un algorithme sous la forme d'un arbre de décision lui permet de décider en contexte de la fonction grammaticale à attribuer au mot et d'en déduire le lemme le plus probable.

- **Brill Tagger** : *Brill Tager* a été créé par Eric Brill dans la cadre de sa thèse en 1993. Cet outil est fondé sur les travaux de Bloomfield et d'Harris. Reposant sur l'idée que l'étude d'une langue peut se fonder sur l'observation de faits linguistiques et indépendamment d'une théorie linguistique particulière, le tagger doit, pour fonctionner, être entraîné sur un corpus de taille restreinte étiqueté manuellement et à partir duquel il infère des règles d'étiquetage (distribution *« extensionnelle »*). Les mots inconnus sont traités à partir d'une hypothèse naïve sur la structure du langage. Enfin, une analyse de la distribution est effectuée afin de réduire les erreurs d'étiquetage.

- le **Cordial Analyseur** : *Cordial Analyseur* est une propriété de la société *Synapse-Développement*. C'est un outil qui intègre entre autres fonctionnalités un étiqueteur morpho-syntaxique. Les résultats de l'étiquetage (texte étiqueté, chiffres absolus, pourcentages, etc.) sont distribués dans différents fichiers en format texte, facilement ré-exploitables dans d'autres cadres d'analyse (méthodes statistiques multidimensionnelles exploratoires, classification automatique, etc.).

- **MBT tagger** : L'étiqueteur MBT est fondé sur un système d'apprentissage combinant deux méthodes d'étiquetages largement utilisées : l'étiquetage stochastique, et l'étiquetage par règles (*Brill tagger*). Un ensemble d'exemples est stocké en mémoire ; chaque exemple contient un mot -ou sa représentation lexicale-, son contexte -antérieur et postérieur- et la catégorie grammaticale à laquelle il est associé dans chaque contexte. Une nouvelle phrase sera analysée de la manière suivante : pour chaque mot de la phrase, le tagger cherchera un exemple d'emploi analogue dans la mémoire et en déduira sa catégorie grammaticale à partir de ses plus proches voisins. Les tags sont donc considérés comme des variables, qui se-

ront assignées aux mots à partir de méthodes de classification. MBT emploie une mesure de similarité qui considère le nombre de tags potentiels qu'il est possible d'associer à chaque mot et qui pondère l'importance de chaque catégorie.

- **LIA Tagg** : LIA Tagg a été développé par F. Béchet au Laboratoire Informatique d'Avignon en 2006. Cet outil tokénise, étiquette, lemmatise et segmente. Il dépend d'une suite d'outils gratuits pour le TAL : le LIA-PHON (convertisseur text-to-phonème), le LIA-TAGG (étiqueteur / lemmatiseur), LIA-SCT (arbre de classification sémantique, LIA-NE (reconnaissance d'Entités Nommées) et LUNAVIZ (outil de visualisation de lattice -treillis-).

- **Stanford POS Tagger** de l'Université Stanford. Il inclut des modules en anglais, arabe, chinois et allemand. Il parcourt le texte dans une langue et affecte les parties du discours pour chaque mot (ou *token*) comme les noms, les verbes, les adjectifs, etc.

- **Apache Open NLP** : est un logiciel maintenu par la Fondation Apache. Il se charge en particulier des tâches d'étiquetage grammatical et d'extraction d'entités nommées. Utilisé avec des textes bruts (exemple : pages Web), il nécessite cependant un nettoyage soigné du texte.

- **Xelda** : *Xelda* est un analyseur morpho-syntaxique ; il utilise la technologie linguistique XFST (technologie des automates à états finis) et a été développée par *Xerox*.

Ce sont les principaux outils de traitements morphologiques. On développons dans la partie Expérimentation le dernier exemple cité à savoir **Xelda**.

3.3.2.2 Le niveau syntaxique

Le niveau syntaxique concerne l'agencement et les relations structurelles des mots dans un énoncé. C'est le niveau conceptuel où l'articulation des séquences de mots et de séquences grammaticales s'établissent afin de valider leur formation. Les énoncés en LN ne sont pas simplement des suites de mots, mais sont organisés en constituants de taille supérieure *i.e* les syntagmes, qui entretiennent entre eux des relations syntaxiques de dominance et de contrôle.

Ces relations syntaxiques peuvent se représenter de plusieurs façons :

1. **le modèle en constituants** considère des groupes de mots, ou syntagmes, généralement centrés sur la tête d'un syntagme (*head*) : le verbe, le nom, l'adjectif détermine alors le syntagme lui-même : syntagme verbal (SV), syntagme nominal (SN), syntagme adjectival (SA). Ces syntagmes peuvent eux-mêmes être éléments d'autres syntagmes.

2. **le modèle en dépendance** considère directement les mots de tête et leur attache les mots qui en dépendent. Les relations grammaticales (sujet-verbe, verbe-objet, verbe-objet-indirect) permettent de représenter la fonction des groupes de mots les uns par rapport aux autres.

Un but important de l'analyse syntaxique est donc d'identifier les différents constituants et sous-constituants, de repérer les relations que ces groupes entretiennent entre eux ainsi que les fonctions syntaxiques qu'il remplissent (sujet, objet direct, object indirect…).

La prise en compte des syntagmes offre une description plus riche du contenu informationnel et améliore donc le traitement des RI. Bien sûr, leur impact dépend de la qualité des informations extraites et de la façon dont elles ont été analysées par les différents outils présentés ci-dessus. Même si des difficultés subsistent comme des ambiguïtés dues à la morphologie d'un mot présentant une graphie identique mais appartenant à une catégorie grammaticale différente, les techniques de TAL apparaissent comme une valeur ajoutée par rapport aux mécanismes traditionnels.

3.3.2.3 Le niveau sémantique

D'une manière assez générale, la sémantique étudie le sens des énoncés. Elle renvoie à une relation entre des objets de la situation considérée (participants, institutions, lieu, temps) [ADA 90].

Les traitements morphologique et syntaxique ne traitent que d'une partie des ambiguïtés de sens. L'analyse sémantique va s'intéresser aux regroupements de termes synonymes, aux familles de termes, aux réseaux de relations sémantiques entre les termes voire entre les textes et les termes. L'objectif est donc de réduire la polysémie et la synonymie, deux phénomènes fondamentaux dans les applications du TAL notamment lors de traductions automatiques, de générations de paraphrase. Ceci devrait permettre d'augmenter la précision des systèmes.

On distingue généralement deux niveaux d'analyse :

- **Prise en compte des relations sémantiques entre les termes** : il s'agit de traiter des relations de synonymie, d'hyperonomie, d'hyponymie, de méronymie et de liens sémantiques. Les relations de **synonymie** désignent des relations entre des mots ou des expressions de formes différentes dans une même langue et ayant un rapport de proximité de sens ou une signification très proche (Exemple : « habit » => « vêtement »). Les relations d'**hyperonymie** représentent des relations sémantiques hiérarchiques d'un terme à un autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus spécifique. (Exemple : « art» => « peinture »). Les relations d'**hyponymie** représentent des relations sémantiques d'un terme à un autre selon laquelle l'extension du premier est inclus dans l'extension du second. (Exemple : « légumes » => « oignons »). Les relations de **métonymie** représentent des relations sémantique qui consistent à remplacer le terme propre par un autre qui lui est proche ou qui en représente une qualité *e.g.* la marque pour la chose ou l'objet. Exemple : « automobile » => « Mercedes ».

  A ce niveau, l'utilisation de ressources externes comme les thésaurus, les ontologies, les dictionnaires de synonymes, les dictionnaires terminologiques (sens en fonction du contexte), dictionnaires spécialisés, les dictionnaires de noms propres, les graphes de relations sémantiques entre concepts, permettent de rendre compte des relations associatives, d'équivalences, de hiérarchie des termes entre eux. A titre d'exemple, le thésaurus *Wordnet*, développé par [MIL 90], présente la particularité de couvrir la majorité des mots (noms, verbes, adjectifs et adverbes) de la langue anglaise et de rendre compte des relations sémantiques qu'ils entretiennent. L'exploitation de ce thésaurus permet une extension de requêtes en incluant des mots sémantiquement reliés aux concepts de la requête originale : les relations de synonymie ou d'hypernonymie qui les structurent très fréquemment peuvent être décortiquées. Chaque mot est donc associé à un certain nombre de sens différents. Le problème réside dans la sélection du ou des terme(s) adéquat(s) pour l'extension de la requête. *Wordnet* regroupe des ensembles de synonymes, appelés *synset*[13]. Chaque synset comporte la liste des synonymes exprimant un même concept. Entre les synsets eux-mêmes, il peut y avoir des relations de synonymie, d'hypernonymie, de méronymie.... Les techniques utilisées ici pour rendre compte de ces relations marient souvent la linguistique et la statistique. Notam-

[13]Un synset est un ensemble de mots de la même catégorie qui peut se substituer dans un certain contexte. Par exemple, {car, auto, automobile, machine, motocar} forme un synset parce qu'ils peuvent servir à référer au même concept.

ment, beaucoup de travaux portent sur les extensions de requêtes dont le principe est l'ajout ou la précision de la requête de l'utilisateur, en important des connaissances sémantiques issues de différentes analyses. Ces apports peuvent être des termes intrinsèques aux documents à la demande ou extraits d'autres requêtes similaires ou encore résultant d'une analyse sémantique latente (*Latent Semantic Analysis* ou LSA). Cette dernière s'appuie sur une représentation multidimensionnelle de la signification des mots dans la langue.

- Le second niveau d'analyse sémantique est celui de la pragmatique et de l'implicite : il s'agit alors de déchiffrer plusieurs éléments comme les non-dits et les connotations d'un énoncé pour lesquels une connaissance extérieure au texte est nécessaire *e.g.* connaissances des situations décrites, des savoirs abordés, de la société en général ainsi que des facultés et des modélisations propres à l'être humain dont nous ne pouvons pas (encore) rendre compte à l'aide d'une machine.

Les connaissances sémantiques qui sont intégrées au sein des SRI peuvent donc être enrichissantes et efficaces si le traitement de désambiguïsation est adapté. L'ajout de connaissances sémantiques permet de préciser la requête de l'utilisateur en ciblant plus précisément le sens de ses constituants initiaux et les rendant par conséquent moins ambigus. Lors de la phase d'indexation, les informations sémantiques visent à offrir une représentation plus riche des contenus textuels, basés non plus sur leurs graphies mais sur leur sens et leur appartenance éventuelle à un thème. On favorise ainsi un appariement plus fin entre les documents et les requêtes et donc une meilleure réponse au besoin informationnel.

3.4 Traduction du besoin informationnel *via* le langage naturel et le langage de requêtes

Nous avons choisi d'étudier un aspect parfois négligé dans la littérature : l'expression du besoin informationnel à travers plusieurs canaux. Plus particulièrement, nous avons étudié s'il existait des études qui comparaient à la fois l'expression du besoin informationnel via le langage naturel et un autre type de langage (langage codé en base de données, langages de requêtes ou encore langage oral). En effet, dans le contexte actuel, l'apparition de nouveaux besoins et de nouvelles applications liées à la recherche d'informations justifie la pertinence d'étudier plus précisément l'expression du besoin

informationnel de l'utilisateur qui est de plus en plus amené à être exprimé sous de nouvelles formes.

A notre connaissance, peu de travaux proposent en effet de comparer les formulations d'un même besoin informationnel, ceci s'explique par deux raisons principales :

1. premièrement, c'est seulement dans de rares occasions que l'utilisateur est appelé à formuler son besoin informationnel *via* deux types de langage, d'où le petit nombre de corpus accessibles pour traiter ce genre de données ;

2. deuxièmement, il peut exister un décalage -au moins temporel- entre le moment où les deux expressions du besoin informationnel sont exprimées.

Pour autant, ce type d'analyse peut clairement apporter des informations intéressantes pour concevoir des SRI plus efficaces. Certes, la LN ne doit pas forcément être systématiquement utilisée pour exprimer les besoins informationnels, mais son utilisation va nous permettre de mieux appréhender les processus sous-jacents mis en œuvre pendant les phases de formulations.

Lors d'une précédente recherche [LAT 05], nous avions également étudié le passage de la verbalisation des demandes en LN à la saisie d'une requête dans un SRI pour douze sujets : six experts et six novices [14] dans leurs domaines de référence au sein de l'INRIA (Institut National de Recherche en Informatique et en Automatique). Nous avions fait les hypothèses que *(i)* le vocabulaire choisi lors de la saisie de la requête correspondait aux termes employés lors de la formulation de la demande en LN et que *(ii)* l'ordre des termes de la question en LN était identique à celle de la structure de la requête.

Nous avions étudié plus particulièrement les deux phénomènes qui s'opéraient le plus :

1. les modification de l'ordre des énoncés,

2. les modifications des formes de surface.

3.4.0.4 Comparaison de l'ordre des énoncés : langage naturel *vs* langage de requêtes

Parmi les études qui ont analysé l'ordre des énoncés en LN reflétant un besoin informationnel, citons les études menées en interface homme-machine (IHM) *via* le paradigme

[14] Les experts correspondaient à des scientifiques en poste, des chercheurs, des ingénieurs, des doctorants en troisième année et les novices étaient des stagiaires en DEA et des doctorants en première année.

du magicien d'Oz qui consiste à faire croire à l'individu qu'il interagit avec un système alors qu'en réalité il dialogue avec un humain.

Par exemple, la technologie ARTIMIS, utilisée comme support dans les applications qui assistent les utilisateurs dans les tâches de RI en LN, en mode écrit ou vocal. Ces services sont qualifiés d'*intelligents* car sont dotés d'aptitudes cognitives qui leur permettent de développer des stratégies de compréhension en contexte, de coopération, de reconnaissance d'intentions, etc. Cette technologie a été utilisée lors des expériences de [LEG 06] sur les expériences de l'application *Plan Resto* et dans les expériences de [AMI 02] sur *Plan Bourses* .

L'application *PlanResto*, est un prototype de service Grand Public à base de dialogue intelligent en langage naturel oral ou écrit pour la recherche de restaurants à Paris. L'utilisateur peut effectuer une recherche avec trois critères :

1. l'emplacement du restaurant (stations de métro, quartiers, arrondissements, principaux monuments),
2. la gamme de prix,
3. la spécialité culinaire.

Le système lui propose en retour des restaurants correspondant au s'approchant le mieux de sa requête. Une recherche typique avec *PlanResto* se décompose en deux temps :

1. Une phase d'émission des critères, *i.e.* l'utilisateur formule une demande plus ou moins précise en LN et le système l'aide en lui demandant de préciser ses critères,
2. Une phase d'affinage, *i.e.* l'utilisateur a émis ses critères et suivant les réponses du système, il consulte et parcourt les solutions. Il peut aussi demander des informations ou des précisions sur les solutions.

A tout moment, l'utilisateur peut basculer entre les phases 1 et 2. [LEG 06] conclue que l'ordre des informations dans les énoncés utilisateurs ne semblent pas obéir au hasard : un ordre typique pour une recherche de restaurants est en effet exhibé [LEG 05]. L'analyse laisse apparaître que pour une demande sur des restaurants, les critères dans la verbalisation d'une demande d'informations ont bien été arrangés dans un certain ordre :

(1) Spécialité + (2/3) Ambiance/Lieu + (4) Horaire + (5/6) Prix / Service.

Cette étude a révélé que la position de ces critères semble être identique dans une demande d'informations quelque soit le sujet : la *spécialité* (1) , l'*ambiance et lieu* (4) semblent être indiqués en début, le *prix* et les *services plus* (6) se placeraient plutôt en dernières positions alors que l'*ambiance* (2) et l'*horaire* (4) auraient des positions variables, plutôt en début ou en milieu de phrases, lors de l'énonciation de la demande.

Pierrel [PIE 91] a également identifié une structure fixe des informations pour une demande d'horaires ou de tarifs de train. En effet, les informations dans ce cadre obéissent à un ordre donné :

sujet/verbe + verbe infinitif + COD + lieu de départ + lieu d'arrivée + etc

Également, dans [LAT 05], nous avions constaté que l'ordre des informations était resté dans 75 % des cas inchangé : les utilisateurs avaient jugé nécessaire de conserver une certaine logique d'enchainement des informations.

3.4.0.5 Comparaison des modifications des formes de surface : langage naturel *vs* langages de requêtes

Dans [LAT 05], tous les utilisateurs participant à l'expérimentation avaient utilisé une ou deux phrase(s) pour exprimer leur besoins informationnels mais aucun ne l'avait retransmis de façon identique dans le SRI : les sujets et les verbes avaient disparu lors de la saisie de la requête ; ils avaient fait place à une juxtaposition de termes. Nous avions également remarqué que les termes retenus par les sujets pour leurs requêtes étaient spécifiques à leurs domaines d'activités ; il s'agissait pour la plupart de termes techniques qui avaient été utilisés dans la formulation de la demande. Toutefois, il faut préciser que lors du recueil de corpus, les sujets exprimaient dans un premier temps leurs besoins à l'oral ; la construction de la requête aurait peut être été différente si cette étape avait été préalablement établie à l'écrit. En effet, certaines constructions ont pu être modifiées à des fins expressives : de clarté ou de mise en valeur d'une information ou d'une idée. Dans 67% des cas, il n'y avait eu aucun nouveau terme introduit dans la requête. Les catégories grammaticales des termes conservés étaient : le nom, le syntagme adjectivale (« Nom + Adjectif ») et le syntagme nominal de type « N de N » (Nom de Nom).

Pour les 33 % restants,il s'agissait :

- **D'ajouts de termes** : pour préciser leurs thèmes de recherche, des termes avaient été introduits pour orienter la recherche à un niveau plus généraliste,

- **De modifications de termes** : conservation de la catégorie grammaticale *e.g.* remplacement d'un substantif par un autre substantif comme dans « contrôle » => « contrôleur », transformation de la catégorie grammaticale *e.g.* d'un substantif à un adjectif comme dans « bactérie » => « bactérien » ou lemmatisation d'un terme.

A contrario, l'expression des besoins informationnels a été retranscrite de façon identique dans plusieurs cas :

1. Dans le nombre de termes employés lors d'une requête : les deux catégories d'utilisateurs (experts et novices) utilisaient environ le même nombre de termes par requête : quatre pour les novices et cinq pour les experts,

2. Dans l'ordre des termes entre une demande en LN et une requête : aussi bien chez les experts que chez les novices, l'ordre des termes entre l'expression du besoin informationnel en LN et celui dans la requête restait inchangé dans 75 % des cas. Nous avions pu en conclure que l'ordre des informations émis dans un énoncé était difficilement dissociable de l'ordre des critères ainsi restitué dans une requête.

Néanmoins, nous avions pu observé quelques variations dans la façon dont les experts et les novices avaient transformé l'expression de leurs besoins en LN puis en requêtes : les novices étaient plus nombreux (dans 75 % d'entre eux) à avoir ajouté des termes. Inversement les experts (également 75 %) étaient les plus nombreux à avoir modifié des termes. Il est possible qu'un effort ait été fait de la part des experts pour faire correspondre leurs besoins informationnels en une requête compréhensible et adaptée au SRI.

Ce premier travail concernait un nombre réduit d'expériences et était centré sur des utilisateurs de SRI ayant tous une tâche professionnelle et très spécifique à effectuer. La nécessite d'effectuer une expérience sur un plus large corpus (avec des types d'utilisateurs et des tâches de RI différenciés) nous est alors apparu essentielle.

Nom du moteur	url du moteur	Résultats à la question	Commentaires
Bing	www.bing.com	*Bing* donne parmi ses premiers résultats la page *Wikipédia* qui a pour intitulé « President of the Senate » (« Président du Sénat ») mais le nom du président lui-même n'apparait pas car cette page n'est pas spécialement dédiée à la France.	L'apport de la technologie sémantique dans *Bing* reste difficilement interprétable car ce moteur renvoie qu'une liste de liens.
Hakia	www.hakia.com	*Hakia* surligne dans ses résultats l'information adéquate : « President of the French Senate President of the Senate since 2011, Jean-Pierre Bel is second in line in the French governmental hierarchy » (Président du Président du Sénat français du Sénat depuis 2011, Jean-Pierre Bel est en deuxième ligne dans la hiérarchie gouvernementale française).	*Hakia* ne se limite pas à une seule source ; il indexe le web au même titre que les moteurs de recherche traditionnels. Les résultats sont ordonnés en plusieurs parties : (a) les résultats provenant du Web ou de sites de confiance, (b) des images, (c) des news. *Hakia* propose une indexation qui n'est pas basée sur des critères de popularité mais bien à partir de critères sémantiques (développement d'un moteur d'indexation : QDEX).
Ask	www.ask.com	Nous obtenons comme réponse « Gérard Larcher » qui l'était en 2010 alors qu'il s'agit bien de Jean-Pierre Bel depuis 2011.	*Ask* se décline en français. L'utilisateur obtient, en réponse à sa question, plusieurs liens ainsi que des résultats sous forme de Q/R populaires. Les résultats ne sont pas très satisfaisants car (1) la demande en LN doit figurer dans les Q/R pour obtenir une réponse, (2) les informations ne sont pas régulièrement mises à jour.
Google	www.google.fr	Google renvoie en première page des résultats le site du Sénat (www.senat.fr/presidence/) qui mentionne dans le *snippet* Jean-Pierre Bel.	Le leader mondial de la recherche comprend aussi le langage naturel, sans pour autant proposer de réponses directes aux questions formulées. Il propose des résultats dont la réponse n'est pas comprise dans les textes mais dans la liste de résultats proposés.

Figure 3.2 – Comparaison de quelques moteurs de recherche autorisant la LN comme langage d'interrogation

Deuxième partie

Expérimentations

Nous nous intéresserons ici à une étude expérimentale de la formulation du besoin informationnel d'un utilisateur d'un SRI à travers son expression via deux types d'énoncés : le premier en langage naturelle, également appelé *langage libre-*, le second en langage de requêtes, classiquement utilisé dans les formulaires de recherche. Pour rappel, la requête telle que nous la traitons ici est l'expression du besoin informationnel sous forme d'une interrogation par le biais d'un SRI.

L'objectif de cette approche est d'évaluer la façon dont un même utilisateur verbalise son besoin informationnel à travers des deux types d'énoncés : nous avons pour cela analysé dans quelle mesure les deux types d'énoncés conservaient des liens sémantiques (linguistiques et/ou structurels) entre eux.

Nous nous situons dans un contexte applicatif, à savoir des demandes de remboursements des utilisateurs d'un moteur de recherche dédié à des études économiques en français. Nous avons recueilli via ce moteur, les deux types d'énoncés sur 5 années consécutives -de 2002 à 2007-, totalisant un corpus de 1398 demandes en langage naturel et de 3427 requêtes (une demande peut être le fruit de plusieurs requêtes).

Grâce à des règles et à des patrons linguistiques ainsi que des analyses morpho-syntaxiques et des outils statistiques, nous avons schématisé l'énoncé en langage naturel en huit concepts que nous avons ensuite comparé à l'énoncé de type mots-clés.

Nous présentons dans ce chapitre le détail de nos hypothèses, ainsi que nos expérimentations et nos méthodes d'analyses.

PRÉSENTATION DES HYPOTHÈSES

Nous avons choisi de tester dans quelle mesure un même utilisateur verbalise son besoin informationnel à travers deux types d'énoncés : le langage naturel et les requêtes. Nous voulons analyser si et dans quelle mesure ces deux types d'énoncés conservent des liens sémantiques (linguistiques et/ou structurels) entre eux. Pour cela, nous basons notre cadre de travail sur trois hypothèses.

- **H1** : le choix du vocabulaire fait pour la saisie de la requête correspond aux termes employés lors de la formulation de la demande en LN. Nous basons cette hypothèse sur [LAT 05] ou de façon éloignée, les travaux réalisés sur la reformulation de [BRU 97]. Ces auteurs indiquent en effet que la reformulation reste souvent très proche de la requête initiale. Les modifications apportées semblent mineures et consistent souvent en *(a)* la répétition de la requête initiale, *(b)* l'augmentation ou la diminution de quelques mots, *(c)* le changement de l'orthographe de la requête, *(d)* l'utilisation de formes dérivées ou d'abréviations.

- **H2** : l'ordre des termes de la question en LN est identique à la structure de la requête. Nous faisons ici l'hypothèse qu'il existe un ordre typique des informations dans la structure de demande de renseignements. Pour [LEG 05] dont les travaux portaient sur la structure des demandes de renseignements de restaurants, l'ordre des informations dans un énoncé est difficilement dissociable de l'ordre des critères dans la recherche et leur enchaînement ne semblent pas obéir au hasard. Les travaux de [PIE 91] ont également mis en évidence une structure fondamentale dans une demande d'horaires de trains où les informations obéissaient à un ordre donné (sujet / verbe + verbe infinitif + COD + lieu de départ + lieu d'arrivée + ...). [VEI 85] indique enfin que les sujets jugent nécessaire de conserver une organisation des informations. La même structure de la phrase est conservée mais des délimiteurs linguistiques étaient souvent modifiés [VEI 85].

- **H3** : la demande en LN ainsi que la requête conservent des éléments inhérents à la tâche à effectuer par l'utilisateur. Pour [JAN 08], la requête est un élément clé de cette expression d'intention. [BRO 02], [KAN 03], [ROS 04], [JAN 08] ont identifié des types de requêtes en fonction des buts des utilisateurs et ont relevé que les recherches d'informations pouvaient être de nature **informationnelle** (*i.e.* tâches de recherche d'un thème particulier), **navigationnelle** (*i.e.* tâches de recherche

de page d'accueil et de navigation) ou encore **transactionnelle** (*i.e.* tâches de recherche de services). Les requêtes transactionnelles ont pour but d'accéder à un site web afin d'y effectuer une interaction ultérieure comme un achat, un téléchargement ou d'autres utilisations de services proposés par le site. Également, pour [CHA 96], les demandes comportent des indices lexicaux ou syntaxiques permettant en effet de repérer les buts explicités par les locuteurs. Nous faisons ici l'hypothèse que *(a)* notre corpus contient également des éléments linguistiques et structurels reflétant les tâches de la RI des utilisateurs qui *(b)* se retrouvent eux-mêmes présents dans la requête.

CHAPITRE 4

CONTENU DE L'EXPÉRIENCE

Pour la validation de nos hypothèses, nous avons choisi de constituer notre propre corpus. Ce choix se justifie pour plusieurs raisons :

1. La réutilisation de corpus déjà existants n'était pas adaptée aux objectifs que nous nous étions fixés. En effet, les corpus existants proposent des listes de requêtes avec le nombre de termes, le nombre de résultats obtenus et le nombre de pages consultées. Ce type de données ne répond que partiellement à notre problématique. Certains corpus, notamment ceux que l'on pourrait extraire d'un site de Questions-Réponses n'apportent qu'un versant des données nécessaires à notre analyse à savoir l'expression du besoin informationnel en LN.

2. Par ailleurs, les analyses s'effectuent généralement sur des moteurs de recherche généralistes, dont les résultats sont disparates et varient énormément par leur forme et leur contenu. Une étude systématique sur ces corpus aurait donc été problématique.

3. Enfin, le type de profils d'utilisateurs retenu dans la littérature ne forme jamais un tout cohérent : ces utilisateurs pouvant tantôt être désignés comme experts ou novices de la recherche d'informations, tantôt experts ou novices d'un domaine d'expertise, ou encore catégorisés en fonction de leur âge, de leur formation, de leur catégorie socio-professionnelle, etc. A notre connaissance, un corpus présentant un échantillon représentatif d'une expression d'une demande en LN, sa reformulation en requête(s), et ceci différencié en fonction de leur but de RI n'existe pas dans un domaine bien défini et/ou sous une forme exploitable.

Pour le recueil de notre corpus, nous nous situons dans un contexte applicatif spécifique, à savoir un moteur de recherche dédié à des études économiques en français. Nous avons recueilli les demandes en LN et les requêtes en français effectuées sur 5 années consécutives (de 2002 à 2007) totalisant un corpus de 1 398 demandes en LN et de 3 427 requêtes (une demande en LN pouvant être formulée par une ou plusieurs requêtes de la part de l'utilisateur). Ce corpus nous a été gracieusement mis à disposition par la société Ubiquick.

Les objectifs principaux du recueil de ce corpus sont multiples :

1. Nous voulons tout d'abord obtenir une description d'un besoin informationnel qui soit une expression librement exprimée en une seule fois. Dans la plupart des cas, l'utilisation des requêtes des moteurs de recherche constitue le seul moyen d'analyser de larges quantités de données concernant les utilisateurs impliqués dans une tâche de recherche d'informations réelle. Mais ce type de données ne donne à aucun moment accès au but réel de l'utilisateur ou à toute autre information sur son environnement (qui est alors simplement deviné). De ce fait, les buts des utilisateurs ne peuvent être qu'inférés avec plus ou moins de précision, notamment en regard du degré de précision des requêtes.

2. De plus, nous voulons étudier de façon concomitante les demandes en LN et les requêtes associées afin de ne pas oublier le sens et le ou les objectif(s) de la démarche de RI : dans quel(s) but(s) et dans quel(s) contexte(s) ? Nous voulons des résultats comparables et transversaux.

4.1 Présentation du moteur de recherche

`Plusdetudes.com`, le moteur de recherche lancé en 2001 par la société *ubquick*, propose un accès illimité à une base de données de rapports économiques (statistiques et études économiques, études sur les tendances d'un marché, profils d'entreprises...). Le site `Plusdetudes.com` a fermé en 2011 pour s'étendre à l'international ; il s'appelle désormais ReportLinker.com. Le même type de données est proposé mais la langue est différente ; tout est désormais en anglais. La page d'accueil du moteur `Plusdetudes.com` est présentée par la Figure 4.1.

Le fonds documentaire est d'environ 10 000 études de marché, couvrant 450 secteurs d'activités économiques et organisé autour de six axes principaux dans le thésaurus sectoriel : Agroalimentaire, Technologies de l'information et Médias, Biens et services de consommation, Sciences de la vie, Industrie, Services. Les premiers niveaux du thésaurus sont présentés sur la page d'accueil du moteur de recherche à la Figure 4.2.

4.1.1 Le Service Après Vente (SAV) du moteur de recherche

Pour accéder à cette base de données, deux types d'abonnements étaient proposés : *(1)* un accès de 24H à cette base pour 39 euros ou *(2)* un abonnement mensuel pour 55 euros. Une fois le compte client créé et l'abonnement effectué, un accès illimité aux

Figure 4.1 – Page d'accueil du moteur de recherche Plusdetudes.com

Figure 4.2 – Secteurs d'activités du moteur de recherche Plusdetudes.com

documents (recherches, lectures et téléchargements) était mis en place par la société. Les deux types d'abonnement sont présentés par la Figure 4.3.

4.1.2 Les clients du moteur de recherche

C'est dans le cadre de leurs recherches d'informations -professionnelles ou personnelles- que les clients utilisent le moteur de recherche. Quand leurs recherches s'avèrent infruc-

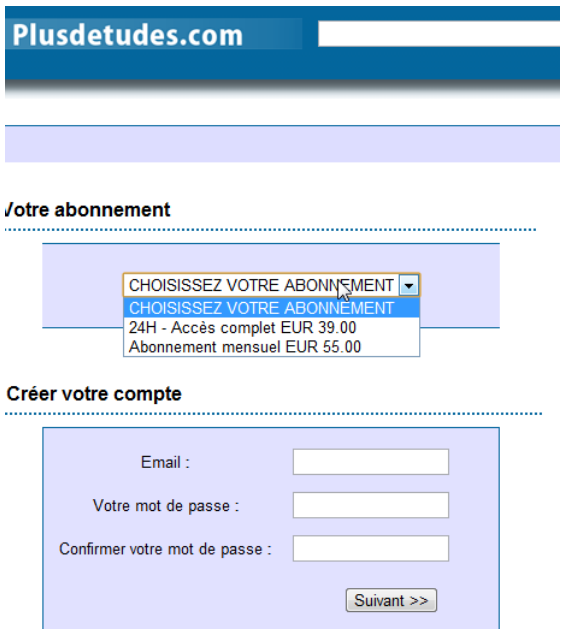

Figure 4.3 – Abonnements du moteur de recherche Plusdetudes.com

tueuses ou tout simplement lorsqu'ils ont du mal à utiliser les fonctionnalités du moteur, certains clients contactent le SAV pour deux types de besoins :

1. **obtenir de l'aide** : formuler leurs requêtes, demander l'avis d'un expert ou pour avoir un accès facilité aux documents.

2. **obtenir un remboursement** : quand les clients ont payé leurs abonnements et qu'ils ne sont pas satisfaits des résultats obtenus.

Ce service client est mis en valeur sur le site ; il suffit de créer un compte pour bénéficier de l'aide d'un des consultants de l'entreprise. Pour cela, les utilisateurs doivent remplir un formulaire SAV qui se présente en plusieurs champs :

1. **les champs « Contacts »** : ces champs permettent d'obtenir des données sur l'identité du client comme le nom, prénom, fonction, nom de l'entreprise.

2. les **champs « Expression de la demande »** : correspond à la formulation libre de la demande. Ce champs est représenté par la Figure 4.5.

A noter, que deux types de démarches peuvent être observées :

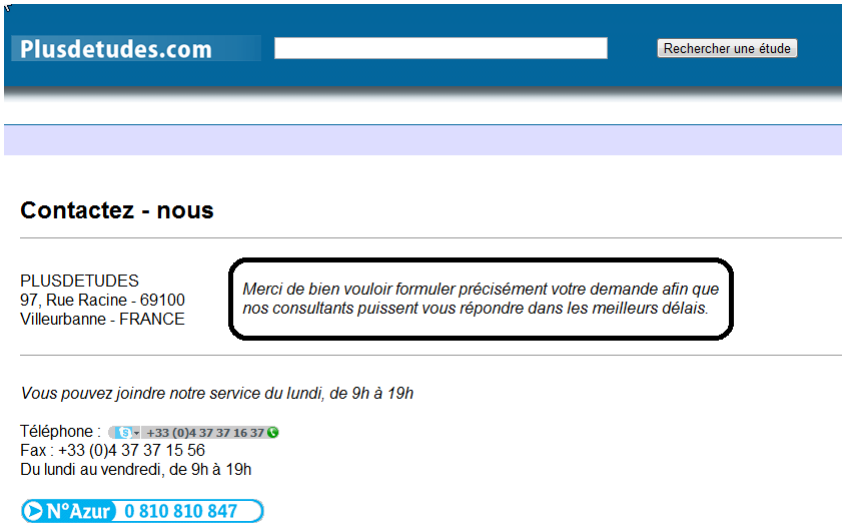

Figure 4.4 – Expression de la demande dans le formulaire SAV

1. la première démarche de renseignement : les utilisateurs ont exécuté une ou des recherches(s) dans la base documentaire (avec consultation ou non de documents) avant de contacter l'entreprise et donc d'expliciter leurs demandes en LN dans le formulaire SAV. Les journaux ou *logs* de requête nous permettent en effet d'identifier les dates et heures des requêtes ; les formulaires SAV contiennent également les dates et heures d'envoi. Ces informations nous permettent donc la comparaison temporelle entre les deux. Voir notamment la section 4.2.1 à ce sujet.

2. la seconde démarche de renseignement : les utilisateurs ont d'abord explicité leurs demandes en LN dans le formulaire SAV avant d'essayer des requêtes dans le moteur de recherche.

Nous choisissons dans le présent corpus d'étudier seulement les requêtes issues du premier procédé ; trop peu de démarches représentent le second probablement parce que les internautes ont instinctivement envie d'essayer une ou plusieurs requête(s) avant d'avoir recours au consultant de l'entreprise. Nos résultats devront donc en tenir compte dans l'interprétation des données.

4.2 Recueil des données

Plusieurs types de données ont été recueillis :

1. les journaux ou *logs* de requêtes,
2. les données personnelles des utilisateurs *via* le champ « Contacts » du formulaire SAV,
3. les demandes en LN *via* le champ « Expression de la demande » du formulaire SAV,
4. la ou les requête(s) effectuée(s) sur le moteur de recherche `Plusdetudes.com`,
5. le nombre de documents renvoyés par requête effectuée.

4.2.1 Recueil des *logs* de requêtes

Dans le cadre de la recherche d'informations sur le Web, les journaux ou *logs* de requêtes *« search logs »* correspondent à un fichier journal au format texte contenant les échanges communicatifs entre un système et ses utilisateurs. Ces fichiers enregistrent un certain nombre d'informations comme l'adresse IP, le fichier atteint, la date et l'heure de connexion. Chaque ligne représente un accès à l'une des pages du site internet. Ils sont également définis par [JAN 06] : p.408 comme « un enregistrement électronique des interactions qui sont intervenues durant un épisode de recherche entre un moteur de recherche sur le Web et les utilisateurs cherchant des informations sur ce moteur de recherche ». Ces *logs* de requêtes contiennent également toutes les requêtes soumises par les utilisateurs au moteur de recherche, la date et l'heure de soumission de la requête, les termes de la requête, le nombre et la liste des résultats obtenus. Il est maintenant établi que l'exploitation de ces *logs* peut aider à l'analyse de grandes bases de données [KHO 09] ; ces *logs* fournissent des données d'une grande diversité permettant de refléter la majorité des comportements et des intentions de recherche.

D'après ces *logs* de requêtes, un identifiant utilisateur (ID-CLIENT) est créé, permettant de relier ces *logs* aux formulaires SAV et par conséquent à l'expression de la demande en LN par l'identifiant demande (ID-DEMANDE) et aux données clients par l'identifiant client (ID-CLIENT). La Figure 4.5 représente les liens entre les différents identifiants.

A noter que pour travailler sur ce corpus, les *logs* de requêtes ont été anonymisées afin de conserver la confidentialité des clients du moteur de recherche.

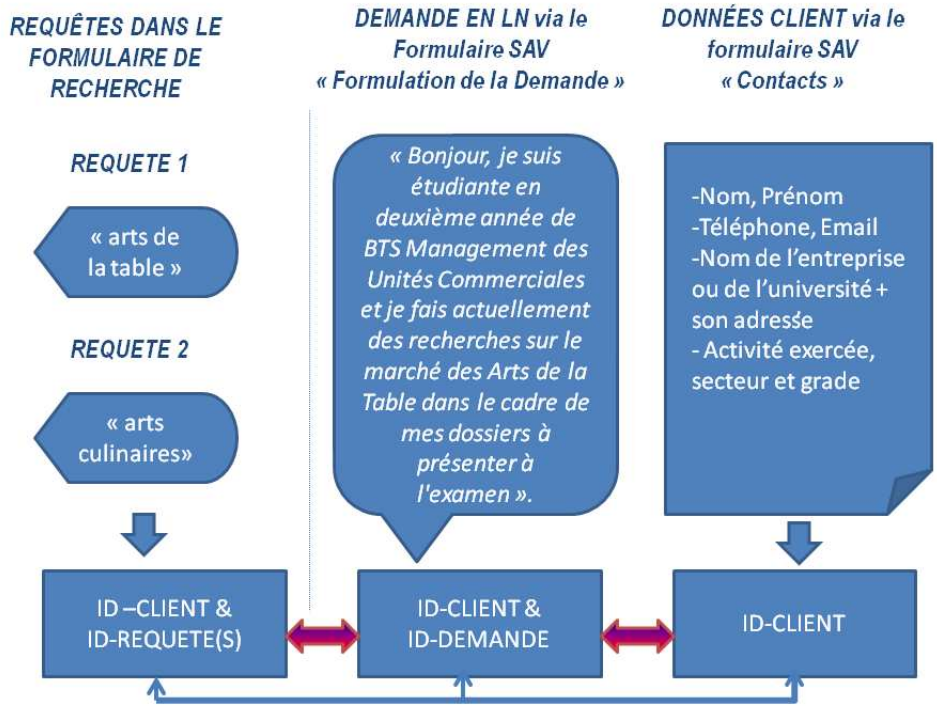

Figure 4.5 – Recueil des données clients par les formulaires de recherche et les formulaires SAV

4.2.2 Recueil des données utilisateurs

Grâce à 1398 formulaires SAV présentés à la section 4.2, nous avons pu obtenir des données assez précises sur les profils utilisateurs. Ces données sont représentées par le tableau 4.1 ; elles concernent l'identité de la personne, l'entreprise (ou université) de laquelle il ou elle dépend ainsi que la fonction exercée.

Identité	Entreprise	Fonction
Nom de l'utilisateur	Nom de l'entreprise	Activité exercée
Prénom de l'utilisateur	Secteur d'activité de l'entreprise	Secteur
Email et téléphone	Adresse de l'entreprise	Grade

Tableau 4.1 – Les données personnelles recueillies via la champs « Contacts »

Ces données seront utilisées notamment pour identifier la tâche de RI à réaliser, développée à la section 5.1.3.

4.2.3 Recueil des demandes en LN

C'est à partir des champs « Expression de la demande » du formulaire SAV présenté à la section 4.1.2, que nous avons obtenu l'expression libre des demandes. Un des avantages de ce corpus est que l'expression du besoin informationnel se fait de manière spontanée (ou au moins, non préparée) afin de s'assurer du caractère « naturel » des résultats. Les utilisateurs s'expriment librement et avec leur propre vocabulaire. Ils n'ont pas à connaître ni à maitriser les techniques documentaires de recherche d'informations. Un deuxième avantage est que l'expression des demandes en LN n'est pas limitée en terme de longueur (elle est non succincte) : l'utilisateur peut exprimer son besoin en plusieurs phrases, sans contrainte de limites de caractères nous permettant ainsi de mieux contextualiser le but de la RI et le profil utilisateur.

4.2.4 Recueil des requêtes

Les requêtes constituent une donnée de base qui, exploitées seules, sont assez équivoques ; elles doivent en effet être mises en lumière avec d'autres informations et ressources afin de pouvoir faire émerger -le cas échéant- des corrélations entre les caractéristiques des requêtes et les paramètres contextuels de la recherche.

CHAPITRE 5

MÉTHODES D'ANALYSES UTILISÉES

Pour rappel, nous avons pour objectifs d'analyser :

1. les expressions des besoins informationnels en LN : en identifiant l'objet même de la demande (le « référant ») ainsi qu'en relevant d'autres types de données comme la couverture géographique, le budget, le scope temporel...

2. les requêtes effectuées dans le moteur de recherche `Plusdetudes.com` : le nombre et l'ordre des termes par requête, la langue utilisée, la catégorie grammaticale des termes, le genre et le nombre des termes, la présence ou non d'opérateurs logiques.

3. la tâche de RI des utilisateurs : exprimée dans la demande en LN et qui peut être contextualisée par des informations sur l'identité *via* le formulaire « Contact » SAV ;

Nous pouvons supposer que si ces caractéristiques sont fortement liées à un aspect spécifique des demandes et besoins informationnels et de de manière relativement stable en fonction du contexte de la recherche, alors il serait possible de trouver des corrélations entre ces caractéristiques et les buts des utilisateurs dans le contexte considéré. C'est ce que nous permet l'analyse concomitante des formulaires SAV et des requêtes sur le moteur de recherche. Nous pouvons ainsi contextualiser la demande avec plusieurs éléments : le profil utilisateurs, les types de besoins informationnels ainsi que les tâches de RI.

Nous avons développé un environnement nous permettant d'écrire des règles qui prennent en compte toutes les informations disponibles soit dans la demande en LN, soit dans les formulaires SAV, soit dans la requête elle-même. Pour cela, nous avons utilisé l'identifiant client, relié à sa demande en LN et aux recherches effectuées. Nous avons délimité à 12 heures maximum l'écart possible entre les requêtes effectuées sur le moteur de recherche et le recours au SAV. Au delà de cette durée, nous estimons qu'il n'y a pas de lien entre les recherches effectuées et la demande de remboursement[1].

[1] A noter que nous ne sommes pas penchés sur les travaux de détection automatique des différentes sessions de recherche [JAN 07] [LEV 13a], [LEV 13b] ; en effet, une analyse humaine nous a permis d'identifier que très peu de sessions de recherche étaient multi-tâches (*i.e.* un utilisateur effectuant plu-

L'enjeu de ces différentes analyses est de comparer le besoin informationnel exprimé en LN et en langage de requêtes en fonction des informations conservées, supprimées, modifiées selon que l'utilisateur s'exprime *via* un formulaire SAV en LN ou *via* une barre de recherche avec des requêtes.

Pour cela, nous présenterons en détail dans les sections suivantes :

- la chaîne d'analyse pour le traitement des demandes en LN,
- la chaîne d'analyse pour le traitement des requêtes,
- les différents outils utilisés.

5.1 Chaîne d'analyse d'une demande en LN

L'objectif de cette analyse est d'appréhender la structure des demandes en LN (comment sont-elles formulées) mais aussi le nombre et le type d'informations (comment est exprimé le secteur d'activité recherché ? L'utilisateur précise t-il d'autres informations comme la zone géographique recherchée ?)

Pour appréhender la totalité de la demande en LN, nous avons développé une chaîne de traitement en quatre étapes :

1. Phase de segmentation des demandes en LN en blocs d'informations ;
2. Phase d'analyse morpho-syntaxique du « référent » des demandes en LN ;
3. Phase d'analyse linguistique de certains traits morphologiques, syntaxiques et sémantiques des blocs d'informations ;
4. Phase de distinction des demandes en LN par types de tâches de RI à effectuer.

Nous expliciterons ces différentes phases dans les sections suivantes.

---

sieurs recherches de nature différente) ; une session utilisateur semble plutôt se déterminer par rapport à un besoin informationnel unique. Nous nous sommes appuyés sur les travaux de [JAN 07] qui définissent une session comme « a series of interactions by the user toward adressing a single information need »[2] ou encore ceux de [SIL 99] qui indiquent : « a session is meant to capture a user's attempt to fill a single information need. »[3]

5.1.1 Phase de segmentation des demandes en LN en blocs d'informations

Nous nous concentrons dans cette phase de segmentation sur un scénario de recherche particulier (*i.e.* la recherche de données économiques via un moteur de recherche en français). Basé sur l'analyse des formulaires SAV, on constate que la plupart de ces demandes comporte une structure sous-jacente. Par conséquent, un ensemble de règles a été écrit pour décrire les différents scénario de la structure de la demande en LN. Cette dernière s'articule autour de plusieurs éléments, désignés comme autant de *blocs d'informations* [BON 79] :

- blocs d'informations de salutations (début) [SALUTATION-DEBUT] : « Bonjour », « Madame »,« Monsieur »,...
- bloc d'informations présentant la fonction [FONCTION-CLIENT] : l'utilisateur a tendance à expliquer si sa démarche de recherche d'informations s'inscrit dans un processus professionnel et/ou personnel ; cette présentation est notamment très présente chez les étudiants ou les créateurs d'entreprises ; « je suis responsable chez WWF Vacances », « je suis étudiante », « je travaille actuellement chez » ...
- bloc d'informations annonçant le contexte de la recherche [CONTEXTE] : il peut s'agir du contexte professionnel, personnel ou universitaire dans lequel vient s'inscrire le besoin informationnel : « dans le cadre de mon stage de fin d'études », « dans le cadre d'une étude sur la qualité des approvisionnements dans le secteur de », « en vue de la création de ma future entreprise »...
- bloc d'informations introduisant une intention de recherche [INTENTION -RECHERCHE] : il s'agit du sujet / verbe annonçant l'objet de la recherche : « je souhaiterais obtenir des données sur », « nous voudrions des informations sur », « je recherche »...
- bloc d'informations indiquant le type de données recherchées [TYPES-DONNEES] : un nombre assez important de demandes en LN mentionne le type d'informations économiques recherchés comme un business plan, le chiffre d'affaires, le nombre de ventes, la liste des fusions et acquisitions, les acteurs de la chaîne de commercialisation, etc.). Par exemple, nous avons rencontré : « les parts de marchés de », « le chiffre d'affaire de », « un business plan de », « la liste des concurrents de »...

- bloc d'informations définissant le ou les référent(s) [REFERENT][MAR 76] : *i.e.* la verbalisation de l'objet même de la recherche, ce sur quoi porte le but de la démarche de RI. Exemples de référents : « revêtement de sol », « pierre naturelle », « télésurveillance », « lavage auto », « Groupe Coca-Cola » ; « produits laitiers », « boissons rafraichissantes non alcoolisées », « arts de la table », « Coca-Cola »…
- bloc d'informations donnant des précisions [PRECISIONS] : il peut s'agir d'exemples ou des détails concernant les types de données. Exemples : « type Häagen Dazs », « et plus particulièrement de la pyrotechnie et des pyromécanismes à usage civil »…
- bloc d'informations de salutations (fin) [SALUTATION-FIN] : les formules classiques de salutations à savoir « avec mes remerciements », « bien cordialement »…) peuvent être précédées d'une demande d'aide comme « auriez-vous ces informations ? » , « Comment pouvez vous m'aider ? », « Est-ce possible ? ». Nous les avons alors inclus dans ce même bloc.

Exemple :

> « Bonjour. Je suis étudiant en Marketing et je voudrais des informations sur l'immobilier (comme le chiffre d'affaire par exemple). Merci par avance. »

Se traduisant par :

> « Bonjour [SALUTATION-DEBUT]. Je suis étudiant en Marketing [FONCTION-CLIENT] et dans le cadre d'une étude de cas [CONTEXTE], je voudrais des informations sur [INTENTION-RECHERCHE] l'immobilier [REFERENT] (comme le chiffre d'affaire par exemple) [PRECISIONS]. Merci par avance [SALUTATION-FIN]. »

Ces blocs d'informations ne sont pas tous remplis par les utilisateurs ; certains n'utilisant qu'un schéma simple de demande d'informations :

> « Je voudrais des informations sur l'immobilier ».

Se traduisant par :

> « Je voudrais [INTENTION-RECHERCHE] des informations [TYPES-DONNEES] sur l'immobilier » [REFERENT].

Des blocs d'information comme celui [CONTEXTE] peuvent se retrouver à trois emplacements possibles dans la demande en LN : [CONTEXTE-1] ; [CONTEXTE-2] ; [CONTEXTE-3] comme présentés dans la Figure 5.1. Ces trois blocs [CONTEXTE] peuvent être utilisés de façon scindée (*i.e* les informations contextuelles sont alors éparses dans ces trois blocs) soit utilisées de façon unique (*e.g.* les informations contextuelles sont données dans le bloc [CONTEXTE-2]).

Exemple de demande en LN où les informations contextuelles se retrouvent à trois emplacements :

> « Étudiante en école de commerce, je dois réaliser une étude de marché. Je souhaiterais avoir des renseignements sur le luxe. Le projet est pour dans 2 semaines. J'ai choisi de cibler mon étude sur les produits Hermès. Malheureusement en tant qu'étudiante, je n'ai pas de budget. Merci par avance. »

Se traduisant par :

> « Étudiante en école de commerce [FONCTION-CLIENT], je dois réaliser une étude sur un secteur [CONTEXTE-1]. Je souhaiterais avoir des renseignements sur le [INTENTION-RECHERCHE] luxe [REFERENT]. Le projet est pour dans 2 semaines [CONTEXTE-2]. J'ai choisi de cibler mon étude sur les produits Hermès [PRECISIONS]. Malheureusement en tant qu'étudiante, je n'ai pas de budget [CONTEXTE-3]. Merci par avance. [SALUTATION-FIN] »

Exemple de demande en LN où les informations contextuelles sont utilisées de façon unique dans le bloc [CONTEXTE-2] :

> « Je suis à la recherche du nombre d'accident automobile pour trouver le nombre de roues en alliages qui on été remplacer ou réparer dans une année complète. Je suis résident du Québec au Canada. Merci de bien vouloir m'aider. »

Se traduisant par :

> « Je suis à la recherche [INTENTION-RECHERCHE] du nombre d' [TYPES-DONNEES] accident automobile [REFERENT] pour trouver le nombre de roues en alliages qui on été remplacer ou réparer dans une année complète [PRECISIONS]. Je suis résident du Québec au Canada [CONTEXTE-2]. Merci de bien vouloir m'aider. [SALUTATION-FIN] »

D'autres concepts associés comme le prix, la zone géographique, les dates ou encore le caractère urgent (le délai) de la demande peuvent apparaître de façon disparates dans les différents blocs d'informations. Nous les avons donc relevés dans des champs différents (cadre de droite dans la Figure 5.1). Pour ce type d'informations, il faut alors pouvoir les extraire quelque soit leurs places au sein des blocs d'information.

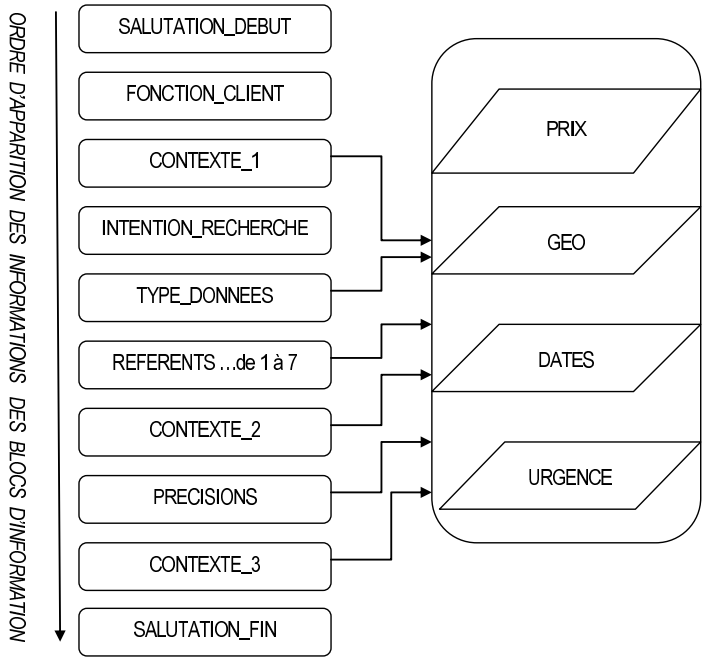

Figure 5.1 – Blocs d'informations pour une demande d'informations en LN

5.1.1.1 Segmentation en blocs d'informations *via* un analyseur linguistique

Pour l'analyse des demandes en LN, nous ne nous livrons pas à une étude quantitative avec l'étude de la totalité de l'expression lexicale des 1398 demandes en LN ; le dépouillement aurait rendu l'entreprise trop longue et difficile dans un premier temps de notre travail. Il s'agit plutôt d'utiliser les régularités que manifeste notre corpus pour

effectuer des découpages et des structurations. Nous nous plaçons sur un domaine applicatif bien spécifique ; même si les formulations sont hétérogènes, elles doivent être probablement conçus sur des structures linguistiques qui peuvent être regroupées (voir les principes de sous-langages ou langages de spécialités développés dans la sous-section **??** à ce sujet). Par exemple, sont répertoriés dans le même bloc : « je voudrais des informations sur [. . .] », « je désire toutes les données concernant [. . .] ».

La démarche ici consiste à dégager peu à peu, à partir d'un corpus d'apprentissage, les régularités à l'œuvre et mettre au point des procédures qui s'appuient sur les principaux motifs séquentiels fréquents (un pronom («je») suivi d'un verbe («vouloir», « désirer ») suivi d'« informations » ou de « données ».

5.1.1.2 Extraction de concepts associés : les Entités Nommées

En parallèle de la segmentation de la demande en blocs d'informations, sont extraits plusieurs concepts importants grâce à la technologie des *cartouches de connaissances* : la zone géographique [PROPER-NAME-COUNTRY], le nom des entreprises et/ou de marques [PROPER-NAME-BUS], les expressions de temps [NUM-DATE], de quantités, pourcentages [NUM-QUANTITY], les valeurs monétaires [NUM-PRICE]. Cette activité de recherche regroupe ce que l'on appelle la recherche d'Entités Nommées (EN) - également désignée en anglais sous *Named Entity Recognition* (NER)- et qui connaît un essor tout particulier ces dernières années avec la recrudescence des données non structurées [CHE 12] [GUO 09], [NAD 07]. Les EN sont alors désignées comme des éléments atomiques dans le texte appartenant aux catégories énumérées ci-dessus. Pour extraire ces entités nommées, la technologie des *cartouches de connaissances* combine à la fois le recours à du vocabulaire sous forme de lexique et à la fois à des règles linguistiques qu'il faut programmer en fonction du contexte. Nous détaillons ci-dessous les types de procédés mis en place :

- la zone géographique [PROPER-NAME-COUNTRY] : appel à un lexique issu du thésaurus géographique interne à la société. Ce thésaurus regroupe à la fois les noms propres des noms de pays (« France ») ainsi que leurs formes adjectivales (« français » dans « marché français »),

- les noms de marques et/ou d'entreprises [PROPER-NAME-BUS] : appel à la fois à un lexique comportant le nom et les marques des plus grandes entreprises et également appel à des règles linguistiques pour détecter celles qui ne sont pas

renseignées. Ces règles linguistiques se basent sur plusieurs éléments : l'apparition d'une majuscule qui ne soit pas au début d'une phrase, des mentions de forme juridique des entreprises comme SA, SARL, SCI pour les formes françaises et enfin sur la fonction exercée dans l'entreprise comme « gérant de », « propriétaire de », « secrétaire chez »,

- les expressions de temps [NUM-DATE] : combinaison de règles et de lexiques permettant de recueillir toutes les formes de dates *e.g.* simple mention de l'année, ou mois + année, etc ainsi que des notions de temporalité comme le semestre par exemple,

- les valeurs monétaires [NUM-PRICE] : combinaison de règles et de lexiques permettant la reconnaissance d'entités numériques suivies ou précédées d'un signe monétaire,

- les quantités et les pourcentages [NUM-QUANTITY] : combinaison de règles et de lexiques basés sur les nombreux cardinaux et certains symboles mathématiques comme le pourcentage.

Le délai de la demande et les *tokens* inconnus sont également pris en charge dans la *cartouche des connaissances*. Un *token* désigne une entité (ou unité) lexicale, dans le cadre de l'analyse lexicale.

A noter que les résultats obtenus dépendent de la qualité des lexiques et des règles linguistiques mises en place. Dans [LAT13b], les résultats obtenus étaient très satisfaisants. Nous avons évalué manuellement les résultats obtenus sur 200 demandes en LN. Il s'avère que la reconnaissance géographique est assez pertinente : moins de 2 % de bruit. Les données numériques et les format dates sont également bien reconnus : moins d' un (1) % de bruit. La reconnaissance des noms de marque et/ou de noms d'entreprise s'est avérée plus fastidieuse puisque celle-ci s'élève à 7 % de bruit (« Wilson »,« Décathlon »). Ceci s'explique par deux faits : le premier si ces noms de marques ou d'entreprises ne comportent pas d'entités d'entreprise comme SA, SARL, Inc, Ltd. etc. Le second si ces noms de marques ou d'entreprises ne sont pas annoncés par certains éléments déclencheurs comme des fonctions « PDG » de Wilson, des acteurs comme « concurrents », « leaders » de Wilson, etc.

5.1.1.3 Extraction des motifs séquentiels fréquents

La recherche de motifs fréquents a fait l'objet de nombreux travaux en fouille de données. Cette approche consiste consiste à analyser et à extraire des enchaînements de séquences apparaissant ensemble avec une fréquence significative. Elle a été introduite par [AGR 95]. L'extraction de motifs séquentiels se fait à partir d'une liste ordonnée (la séquence) d'ensembles appelés *items*. Un ensemble d'*items* est communément appelé *item-set* et noté $i_1, i_2 \ldots i_m$ où les i_j sont des *items*. Une séquence est pour sa part une liste ordonnée d'*item-sets*, notée $\langle S_1 \ldots S_n \rangle$ où les s_j sont des *item-sets*. Il s'agit alors de repérer, dans un ensemble de séquences, des enchaînements d'*items* ayant une fréquence d'apparition supérieure à un seuil donné.

Dans un premier temps, notre méthodologie a consisté à extraire de manière empirique les caractéristiques et les structurations des *item-sets* d'un corpus d'apprentissage constitué d'environ 15 % de notre corpus total soir 200 demandes en LN. Pour cela, nous avons relevé manuellement :

- le vocabulaire et les formulations d'une demande pour les différents blocs d'informations en s'appuyant sur des **marqueurs sémantiques** (« par exemple », « aussi », « concrètement », « en effet »…) ;
- des **marqueurs typographiques** comme la ponctuation (virgule, le point virgule) et les majuscules).

Dans un second temps, nous avons procédé à l'automatisation de cette segmentation. Pour notre étude, cette automatisation s'est effectuée grâce à la technologie *Skill Cartridge*^{TM}(également appelé *Cartouches de Connaissances*) de l'outil *Luxid* [4]. Il s'agit d'une solution commerciale, conçue pour rechercher et extraire des informations dans des données non structurées à partir de règles linguistiques et de lexique. L'utilisation de cette technologie, déjà présente au sein de la société Ubiquick pour d'autres ressources, nous a permis de normaliser la demande d'informations à partir de données non structurées [LAT 13b].

Les règles linguistiques permettent d'ordonner l'apparition des différents blocs d'informations qui sont transformés *via* les *Cartouches de Connaissances* en *concepts*. Ces concepts reposent sur du lexique mais aussi sur des règles de casse, de lemmatisation, de graphie etc. qui permettent soit de faire une correspondance exacte (case-sensitive=0)

[4] http ://www.temis.com

soit de faire une correspondance partielle (basé sur le lemme du terme). L'enrichissement du lexique se fait par apprentissage, tout comme la délimitation des concepts. Certaines séquences constituent en effet des « frontières élastiques », c'est-à-dire qu'elles peuvent tantôt délimiter un concept tantôt en faire partie. La solution réside là encore dans l'apprentissage *ad hoc* de ce corpus bien spécifique. Un premier passage sur le texte relève tous ses concepts. Un second passage les trie et les répartit en groupes. Nous pouvons également délimiter certains *item-sets* que nous voulons voir apparaitre dans les motifs (les contraintes d'appartenance) ou introduire une contrainte dite de « gap » [DON 07] autorisant l'extraction de motifs ne contenant pas nécessairement des *item-sets* consécutifs ou encore des contraintes de début de motif. Par exemple : on pourrait vouloir que le concept [SALUTATION-DEBUT] débute le motif obligatoirement par le terme « bonjour ».

Pour la demande en LN suivante :

> *« Bonjour, je suis étudiant et je recherche des données (comme le chiffre d'affaire) sur le groupe Coca-Cola. Auriez-vous cela ? Merci d'avance. »*

La transformation *via* la cartouche de connaissance va être la suivante :

> *« Bonjour* [SALUTATION-DEBUT], *je suis étudiant* [FONCTION-CLIENT] *et je recherche* [INTENTION-RECHERCHE] *des données (comme le chiffre d'affaire)* [TYPE-DONNEES] *sur le groupe Coca-Cola* [REFERENT]. *Auriez-vous cela ? Merci d'avance.* [SALUTATION-FIN] »

L'évaluation manuelle de cette méthode a montré que dans 75 % des cas le découpage en motifs obtenu était correcte ; 20 % des séquences restaient non étiquetées (*i.e.* pas d'appartenance aux différents blocs d'informations identifiés) et 5 % des séquences étaient étiquetées de manière erronée. Si l'on effectue un tri supplémentaire sur ces 5%, manuel cette fois, on peut distinguer, plusieurs cas de figures :

- mauvais étiquetages, principalement sur les noms propres (problèmes de majuscules),
- mauvaises structurations de l'utilisateur probablement dues à des erreurs lors des phases de corrections orthographiques et typographiques de certains termes.

Les séquences étiquetées de manière erronée ou non étiquetées ont été réaffectées manuellement.

L'avantage de cette approche est qu'elle ne nécessite ni corpus annoté manuellement ni analyse morpho-syntaxique. Elle va nous permettre :

1. de conserver l'ordre des énoncés de la demande ;
2. de travailler de façon transversale sur certaines parties du discours (comme les référents) ;
3. d'automatiser le processus afin de prévenir d'éventuels erreurs de catégorisation manuelle lors de la répartition des données dans les différents concepts (même si ces erreurs peuvent être peu nombreuses) ;
4. de mettre en place une méthode qui puisse être exploitée assez rapidement sur d'autres corpus et donc à plus grande échelle.

Néanmoins, cette méthode non-supervisée nécessite une sélection des motifs pertinents car le nombre de motifs extraits peut être assez important surtout en fonction du corpus test. Cette méthode permet donc de générer beaucoup de modèles qui doivent être traités en fonction de leur redondance et des variations parfois minimes qui sont proposées (on parle alors de *sous-motif*). Pour pallier ce problème, [CEL 10] propose une représentation condensée des motifs qui élimine les redondances entre ces derniers (*i.e. motifs fermés*).

Cette segmentation porte donc sur des formes de sous-catégorisations qui serviront ensuite à formaliser l'expression des demandes en LN et notamment à traiter spécifiquement les différents blocs d'informations (concepts) selon des usages différenciés.

5.1.1.4 Usages différenciés des blocs d'informations

Nous nous sommes orientés vers une approche qualitative avec une analyse différenciée des différents blocs :

- le bloc [REFERENT] nous livre un bon nombre d'informations sur l'objet même de la recherche ;
- les blocs [FONCTION-CLIENT] et [CONTEXTE] contiennent des informations sur le profil utilisateur ;
- les blocs [TYPES-DONNEES], [PRECISIONS] et [REFERENT] nous donnent accès à des concepts associés comme la zone géographique recherchée, la ou (les)

dates demandé(es), les critères de prix s'ils sont présents, ou d'autres éléments de contexte de la recherche comme le délai/l'urgence ou non de la demande.

- le bloc [INTENTION-RECHERCHE] pour savoir comment les utilisateurs verbalisent leurs besoins. Nous avons étudié quelle était la syntaxe générale des phrases : phrases complètes, partielles ou se structurant de la même façon qu'une requête. Nous avons également relevé les temps verbaux, les types de construction des verbes, les pronoms personnels utilisés...
- enfin les [SALUTATION-DEBUT] et [SALUTATION-FIN] ne sont pas étudiés dans ce présent travail ; notre objectif est donc uniquement de les identifier.

5.1.2 Phase d'analyse linguistique des blocs d'informations

Nous avons ensuite analysé ces différents blocs d'informations en deux temps :

1. le premier correspond à une analyse linguistique de certains traits morphologiques, syntaxiques et sémantiques des blocs d'informations de la demande en LN. Ces traits sont énumérés dans le Tableau 5.1 ; ils permettent d'obtenir des informations de type « général » sur la demande elle-même comme sa longueur, son complexité, sa structure grammaticale ou encore le nombre et le type d'entités nommées, etc.
2. le second correspond à une analyse syntaxique plus fine du bloc [REFERENT], considéré dans cette étude comme le pendant de la requête. Les principaux traits relevés pour cette analyse sont spécifiés dans la Figure 6.5 à la page 105. Ils nous permettent d'avoir une vision plus précise sur la composition du [REFERENT] comme le nombre et la structures des termes, etc.

5.1.2.1 Analyse linguistique de certains traits morphologiques, syntaxiques et sémantiques des blocs d'informations

Nous avons eu recours aux caractéristiques linguistiques extraites des requêtes chez Mothe et Tanguy [MOT 05] que nous avons adoptées et notamment présentés dans l'article [LAT 13a]. Ces caractéristiques sont au nombre de seize et correspondent à trois niveaux linguistiques : morphologique, syntaxique et sémantique. Elles sont présentées dans la Figure 5.1.

Traits morphologiques de la demande en LN	
Nombre de mots	NBWORD
Moyenne de la longueur des mots	LENGHT
Moyenne des noms propres (pays)	PROPER_NAME_COUNTRY
Moyenne des noms propres (entreprises, marques)	PROPER_NAME_BRAND
Moyenne des acronymes (acro)	ACRONYM
Moyenne des valeurs numérique (dates)	NUM_DATE
Moyenne des valeurs numérique (quantité)	NUM_QUANTITY
Moyenne des valeurs numérique (prix)	NUM_PRICE
Moyenne des tokens inconnus	UNKNOWN
Traits syntaxiques de la demande en LN	
Moyenne des pronoms personnels	PP
Moyenne des structures syntaxiques correctes	STRUCT_SYNTAX_RIGHT
Moyenne des structures syntaxiques partielles ou incorrectes	STRUCT_SYNTAX_FALSE
Moyenne des phrases interrogatives	INTER_SENTENCE
Moyenne des phrases affirmatives ou négatives	AFFIR_NEG_SENTENCE
Traits sémantiques de la demande en LN	
Moyenne de la valeur polysémique (nbre de fois apparaissant dans le thésaurus)	POLYSEMY
Moyenne des usages des blocs [CONTEXTE] et [PRECISIONS]	CONTEXT
Moyenne de la complexité linguistique (profondeur des nœuds dans le thésaurus)	LINGUISTIC_COMPLEXITY

Tableau 5.1 – Traits linguistiques des demandes en LN inspirés de [MOT 05]

5.1.2.2 Phase d'analyse morpho-syntaxique des « référents » des demandes en LN

Nous avons testé et amélioré la méthode décrite en section 5.1.1 jusqu'à l'obtention d'un ou plusieurs référent(s) pour chaque demande en LN ; quelques rares exemples n'en contiennent pas, nous les avons alors annotés. Nous avons également relevé tous les concepts associés comme la zone géographique, le scope temporel, le type d'informations recherchées, le but de la recherche, etc.

Le nombre de concepts associés au référent pouvant varier de 0 (pas de concept associé, la demande en LN est alors très proche d'une structure requête) à 7.

Exemple n.1 avec 0 [REFERENT] « Je recherche des éléments statistiques »

Se traduisant en concepts par : [INTENTION-RECHERCHE] [TYPE-DONNÉES]

Exemple n.2 avec 1 [REFERENT] « Je souhaiterais avoir des renseignements sur le produit Sunny Delight »

Se traduisant en concepts par : [INTENTION-RECHERCHE] [TYPE-DONNÉES] [REFERENT 1]

> Exemple 3 avec 7 [REFERENTS] « Je souhaite des chiffres clés pour les marchés : santé Europe, beauté Europe, nutrition Europe, environnement Europe, Thalassothérapie, spa Europe, Spor Europe. Sincères remerciements »

Se traduisant en concepts par : [INTENTION-RECHERCHE] [TYPE-DONNÉES] : [REFERENT 1], [REFERENT 2], [REFERENT 3], [REFERENT 4], [REFERENT 5], [REFERENT 6], [REFERENT 7] [SALUTATIONS-FIN].

Nous avons fait ce découpage pour ne conserver que l'expression linguistique des référents. La plupart des éléments présent dans la requête se retrouvent également dans le ou les référent(s). Ce point sera étudié dans la section 6.3.2. Afin de pouvoir être comparé plus finalement avec la requête, ce bloc a fait l'objet d'une analyse qualitative notamment une étude morpho-syntaxique de l'ensemble de ses termes. Pour cela, nous nous sommes à nouveau basés sur les principaux traits relevés par [MOT 05] qui nous permettent d'avoir une vision plus précise sur la composition du [REFERENT].

L'analyseur morpho-syntaxique utilisé est présenté dans la partie 5.3.

5.1.3 Phase de distinction des demandes en LN par les types de tâches de RI à effectuer

Nous superposons aux analyses des demandes en LN déjà existantes une méta-analyse qui concerne la tâche de recherche d'informations à effectuer. Cette tâche de RI déterminera les profils utilisateurs qui nous serviront de base pour l'analyse.

5.1.3.1 Identification de profils d'utilisateurs en fonction des tâches de RI

En parallèle des études distinguant des profils d'utilisateurs en fonction de leur capacité d'expertise (alors identifiés comme *novices* ou *experts* d'un domaine ou des techniques de recherche ou de leurs catégories socio-professionnelles) ; nous nous sommes plutôt intéressés à la question générale du comportement de recherche des utilisateurs et plus spécifiquement sur le lien entre les demandes enregistrées *via* le moteur de recherche et la tâche des utilisateurs. C'est en effet en connaissant le plus précisément possible la façon dont l'utilisateur va élaborer ses stratégies de recherche d'informations qu'il sera possible de lui proposer des outils susceptibles d'améliorer significativement ses recherches, et donc son accès à l'information. Ces connaissances sont complémentaires aux comportements de recherche avec des métriques comme le nombre de documents consultés, la durée de consultation des pages, le nombre de reformulations le cas

échéant, le taux de rebond (aucun document consulté, l'utilisateur ferme sa session de recherche).

L'un des enjeux majeurs de cette analyse est d'adapter les moteurs de recherche de manière à mieux prendre en considération le contexte et les tâches des utilisateurs pour ainsi proposer des résultats plus pertinents en fonction de leur besoin d'informations. Nous proposerons quelques pistes d'améliorations des systèmes dans la partie 7.4. De nombreuses études dont notamment celles de [ING 05], [RAM 06] ou encore [LAU 99] expriment un décalage entre le besoin informationnel et son expression à travers un SRI. Ce décalage peut s'expliquer par notamment plusieurs facteurs : le choix et la combinaison des termes, la longueur de la requête, la langue de l'interface, la méconnaissance des opérateurs de recherche. Pour [ROS 04], le comportement de recherche s'effectue autour de trois axes : *(a)* comment les utilisateurs effectuent une recherche, *(b)* qu'est ce qu'ils recherchent et *(c)* la raison de leur recherche.

Nous avons donc voulu identifier le profils des utilisateurs et ce, à travers deux canaux d'informations :

1. Pour augmenter la « couverture » du profil, nous avons utilisé des informations contenues dans les demandes en LN. En effet, nous avons constaté que la demande en LN pouvait contenir des informations explicites permettant d'enrichir notre connaissance sur l'objectif poursuivi par l'utilisateur. Nous avons eu pour objectif d'identifier les phrases (ou les syntagmes) potentiellement *saillants* pour incrémenter nos résultats. Une phrase est considérée comme *saillante* si elle contient une trace linguistique qui exprime une tâche, un but, un objectif à réaliser, un contexte de la recherche d'informations. Il s'agira d'informations contenues dans le concept [BUT-RECHERCHE] que nous décrivons dans la partie suivante. Un exemple de requête *saillante* serait *acheter une voiture* ; dans laquelle le but est clairement identifié contrairement à une requête qui ne mentionnerait que « voiture » par exemple. La première représente clairement l'intention d'acheter une voiture alors que la seconde peut renvoyer elle-même à un ensemble de buts très variés.

2. le champs « Contacts » du formulaire SAV présenté à la section 4.2.2 pour l'obtention des données personnelles.

Dans notre étude, c'est à partir des analyses concomitantes des formulaires SAV (avec les coordonnées, professions, nom de l'entreprise, demandes explicités en LN) et des

demandes en LN que nous avons pu nous appuyer sur une base d'informations assez solides pour distinguer une typologie de demandes des tâches de RI.

5.1.3.2 Analyse des tâches de RI

Dans un premier temps, nous avons procédé à une analyse manuelle des demandes en LN pour identifier les tâches utilisateur. Si la tâche à effectuer n'a pas été clairement exprimée dans la demande en LN alors celle-ci n'est pas étudiée et est écartée de notre analyse (nous sommes en effet motivés par l'étude de la tâche de RI sous-jacente aux demandes en LN et aux requêtes ; dans le cas contraire, ces demandes et ces requêtes pourraient faire parties du corpus). A la suite de ce traitement, nous avons distingué trois types de besoins informationnels en fonction de la tâche à réaliser :

- le premier est **la création d'entreprise ou le lancement d'un nouveau produit ou encore d'une nouvelle marque** [TACHE-CREA] : les objectifs opérationnels sont alors de se procurer une étude de faisabilité, d'identifier la concurrence éventuelle. Cette tâche peut s'opérer dans un cadre personnel (auto-entrepreneur, en recherche d'un emploi) ou professionnel (développement d'une activité professionnelle ou création d'une nouvelle).

- le second est **la réalisation d'une tâche scolaire** [TACHE-SCO] : les objectifs opérationnels sont alors de préparer un examen, d'écrire un mémoire, de travailler sur une étude de cas... Cela correspond à la réalisation de travaux scolaires ou universitaires.

- le troisième est **l'obtention d'informations dans un cadre professionnel** [TACHE-PRO] : les objectifs opérationnels sont alors de mieux connaitre le marché en ayant des éléments chiffrés, d'identifier les tendances d'un marché ou d'un produit, de faire de la veille stratégique....

La pertinence de la structure de cette typologie se justifie par une frontière assez importante entre les trois différentes tâches de la RI.

La première tâche [TACHE-CREA] correspond à une demande et à des documents bien spécifiques (*i.e.* création d'entreprise ou le lancement d'un nouveau produit) ; cette demande est bien particulière sur des aspects administratifs, financiers et juridiques et représente une part importante du public du moteur `Plusdetudes.com`. C'est en

connaissance de cause, que le moteur avait pré-rempli un champ à cocher pour identifier ce type d'utilisateurs lors de leur inscription. A noter que ce type de tâche peut recouper plusieurs catégories socio-professionnelles (personnes occupant déjà en emploi mais voulant devenir auto-entrepreneur, personnes à la recherche d'un emploi, etc.). Ce type de tâches représente 379 demandes de notre corpus.

La seconde tâche [TACHE-SCO] correspond à la réalisation de travaux scolaires ou universitaires. Il était donc assez aisé de corroborer cette tâche à une catégorie socio-professionnelle à savoir des personnes dont le statut est « Etudiant » (les lycéens et étudiants l'ont ainsi renseigné lors de leur inscription sur le moteur de recherche). Ce type de tâches représente également une part importante du public du moteur de recherche ; 513 demandes de notre corpus.

La troisième tâche [TACHE-PRO] correspond, plus généralement, à l'obtention d'informations dans un contexte professionnel. Nous avons volontairement regroupé les différentes catégories socio-professionnelles sous la tâche commune de recherche d'informations professionnelles ; en effet, plus les utilisateurs ont des profils larges et variés, plus il sera difficile d'adapter les outils de recherche aux documents qu'il convient de retourner. Ce type de tâches représente 506 demandes de notre corpus.
La répartition des demandes par type de tâches est représentée à la Figure 5.2. Nous n'avons pas observé de combinaison de ces tâches.

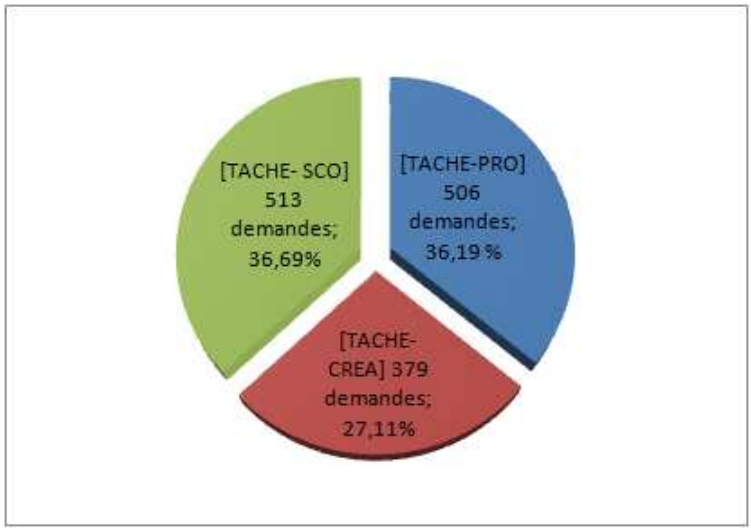

Figure 5.2 – Répartition des demandes en LN selon les types de tâches de RI à effectuer

La tâche à réaliser a donc été identifiée comme le facteur contextuel le plus important dans cette étude.

5.2 Chaîne d'analyse d'une requête : Comparaisons avec le REFERENT

L'analyse de la requête s'est effectuée en plusieurs temps :

1. analyse des traits caractéristiques des requêtes à partir des *logs* de requête(s) : la date et l'heure de connexion, les requêtes saisies dans le moteur de recherche, la date et l'heure de ces requêtes ainsi que le nombre de résultats obtenus. Parmi ces informations, nous exploitons plus particulièrement certaines statistiques comme *(a)* le nombre de requêtes par session utilisateur, *(b)* la longueur des requêtes, *(c)* le nombre de *n-grammes*[5] dans les requêtes ainsi que des spécificités liées aux équations de recherche comme le nombre d'opérateurs booléens utilisées mais aussi le nombre de requêtes formulées en langue étrangère ;

2. analyse plus spécifique sur les entités nommées utilisées dans les requêtes : nom propres de type entité *Business*[6] (nom d'une société, d'une marque), de type entité personne, mais aussi les dates, les lieux géographiques, les entités numériques (notamment les prix).

3. analyse morpho-syntaxique effectuée sur les termes de la requête afin de dégager les catégories morpho-syntaxiques les plus utilisées. Pour cela, nous nous appuyés sur un schéma d'analyse identique et les mêmes traits linguistiques que ceux développés dans la partie 5.1.2.2.

5.2.1 Comparaison [REFERENT] *vs* Requêtes

L'un des intérêts de cette étude est de pouvoir notamment comparer les requêtes et les [REFERENTS] de la demande en LN identifiés lors de la phase de segmentation en

[5]Dans les domaines de la linguistique computationnelle et de la probabilité, un *n-gramme* est une séquence contiguë de *n* éléments d'une séquence donnée de texte ou de la parole. Les items peuvent être phonèmes, des syllabes, des lettres, des mots ou des paires de bases en fonction de l'application. Les *n-grammes* sont généralement collectées à partir d'un texte ou de corpus de conversation. Un *n-gramme* de taille 1 est appelé un « unigramme » ; de taille 2 est un « bi-gramme» ; de taille 3 est un « trigramme », de taille 4 un « quadri-gramme », etc.

[6]Les libellés sont ici en anglais correspondant aux étiquettes en anglais de l'analyseur morpho-syntaxique *Xelda*.

blocs d'informations ceci afin de relever des régularités le cas échéant entre les deux types d'expression du besoin informationnel.

Pour établir cette comparaison, nous nous sommes appuyés sur des travaux déjà existants sur les reformulations de requête : [ANI 03], [TEE 07], [JAN 07], [HE 02] [LAU 99],[BRU 07] et [GUO 08].

Nous avons repris les travaux de [HUA 09] pour concevoir manuellement une grammaire dans le but de générer automatiquement la comparaison des formes de surface (pas d'interprétation sémantique) des [REFERENTS] et des requêtes. Cette grammaire est présentée en Annexe I (page xxiv). Cette grammaire permet de prendre comme point d'entrée de chaine d'analyse le [REFERENT] et d'appliquer des comparaisons exacte de chaine de caractères avec des possibilités de flexibilité (*i.e.* permettre la variation d'un ou deux caractères. Nous avons ainsi comparé les [REFERENTS] et les requêtes sur les points suivants :

1. **Ponctuations et espaces** : nous traitons ici les ponctuations et espaces mais nous les considérons et donc les comptabilisons comme « correspondances totales » (Ex : Auchan, swot => Auchan swot),

2. **Correction orthographique** ou *Spelling Correction* (SC) : une correction orthographique est détectée en utilisant une fonction de distance de Levenshtein [LEI 96]. Cette distance mathématique donne une mesure de similarité entre deux chaînes de caractères (entre la chaine de caractère « AAA » et « AAB », il y a substitution d'un caractère de A en un caractère de B). La distance est d'autant plus grande que le nombre de différences entre les deux chaînes est grand. Cette distance s'applique à deux chaînes de caractères courtes ; elle est donc adéquate pour l'étude de notre corpus car elle permet de faire correspondre le [REFERENT] et la requête. On peut ainsi rapprocher les termes du [REFERENT] et la requête même si ceux-ci présentent des erreurs de frappes, des inversions de caractères ou encore des caractères manquants. Le [REFERENT] et la requête sont classés comme correction orthographique si la distance Levenshtein est égale ou inférieure à deux (exemple : accident automobile => accident automoible),

3. **Ré-ordonnancement de termes** ou *Word Recorder* (WR) : les mots sont réorganisés mais restent inchangés (Ex : Japon Champagne => Champagne Japon). Ici ; le [REFERENT] et la requête contiennent tous les deux les mêmes termes mais inversés.

4. **Suppressions de termes** ou *Remove Word* (RW) : un ou plusieurs terme(s) est/sont supprimé(s) entre le REFERENT et la requête sur la base d'autres termes en commun. (Ex : boulangerie traditionnelle => boulangerie),

5. **Ajouts de termes** ou *Add Words* (AW) : un ou plusieurs terme(s) est/sont ajouté(s) entre le REFERENT et la requête sur la base d'autres termes en commun (Ex : bijouterie => horlogerie bijouterie),

6. **Racinisation** ou stemming (Stem) : cette approche cherche à rassembler les différentes variantes d'un mot autour de son *stem* (*i.e.* une pseudo-racine). Cette procédure traite à la fois des cas relevant de la flexion et de la dérivation. Les techniques utilisées pour procéder à la racinisation reposent généralement sur une liste d'affixes[7] et sur un ensemble de règles de désuffixation construites *a priori*. Elles permettent de retrouver le *stem* d'un mot. L'avantage de ces outils (*stemmer*) réside dans leur simplicité : ces outils gèrent en même temps les morphologies dérivationnelle et flexionnelle. Le stemming est d'ailleurs la méthode la plus utilisée dans les moteurs de recherche [LOU 04]. Les règles du stemming sont décrites à partir de l'algorithme de Porter [POR 80] pour l'anglais ; elles consistent en sept phases successives et une cinquantaine de règles applicables. L'algorithme de Porter a été adapté en français par [PAT 02] avec l'algorithme de désuffixation Carry. Tout comme l'algorithme de Porter, Carry se décompose en phases successives. C'est également le suffixe le plus long qui détermine la règle à appliquer. Ainsi, des demandes portant sur des études de marché sur des «chiennes de race», le mot « chiennes » deviendra « chienn » par suppression du « s », du « e » final, puis « chien » par suppression de la double consonne finale. (Ex : chiennes de race => chien de race),

7. **Acronymes** ou *Form Acronym* (FA) : une transformation sous forme de sigle se produit lorsque la requête est un acronyme (exemple : « SC ») formé à partir des mots du [REFERENT] (exemple : « télévision »). Dans le même registre, une extension de l'acronyme ou *Expand Acronynm* (EA) se produit lorsque le [REFERENT] est un acronyme (exemple : « LLD ») et que la requête contient les termes qui le composent (exemple : « location longue durée »),

8. **Segments de chaînes** ou *Substring* (Sub) : un segment de chaînes *Substring* est définie comme une instance où la requête est un préfixe ou un suffixe strict du

[7] Un affixe est un morphème en théorie lié qui s'adjoint au radical ou au lexème d'un mot.

[REFERENT]. Contrairement à la définition traditionnelle du segment de chaîne, cela ne comprend pas les cas où seuls les caractères à l'intérieur de la première requête sont extraits (Ex : étude sur l'ésotérisme => étude sur l'éso). Dans le même registre, *supertring* (Super) est défini comme une instance où la requête contient le [REFERENT] comme préfixe ou suffixe. Ex : étude sur la para => étude sur la parapharmacie,

9. **Abréviations** ou *Abbreviation* (Abbr) : une reformulation de l'abréviation est effectuée quand les termes du [REFERENT] et de la requête sont des préfixes de l'autre. Cela diffère des segments de chaîne *Substring* et *supertring* qui considèrent les suffixes et préfixes seulement. La reformulation de l'abréviation peut être détectée sur la totalité des requêtes (Ex : marché des déodorants parfumés => marché des déo parfumées),

10. **Substitutions de termes** ou *word substitution* (WS) : le remplacement d'un mot se produit lorsque un ou plusieurs mots dans le [REFERENT] sont remplacés par les mots sémantiquement liés, déterminées à partir des thésaurus internes de la société (présentés à la partie 5.3. Deux termes sont liés si l'un a une relations sémantique (synonyme, hyponyme, hyperonyme, méronyme) avec l'autre. Cette règle est mise en œuvre en deux étapes : soit s'effectuer sur une partie seulement des termes soit sur la totalité des termes (Ex : pain => baguette, hôtel => Kyriad ou encore fruits => banane)

Cette grammaire nous permet donc d'obtenir une comparaison rapide sur les formes de surface entre les REFERENTS et les requêtes. Nous pouvons ainsi répertorier les variations partielles comme celles liées à la ponctuation et espaces, appréhender les ajouts, suppressions ou modifications de termes (sur tout ou une partie de la chaîne de caractères), étudier l'usage de formes développées contre l'usage de sigles, etc. Les comparaisons entre les REFERENTS et les requêtes seront ainsi simplifiées.

5.3 Les outils utilisés

Deux principaux outils ont été utilisés : des ressources externes (thésaurus) et un analyseur morpho-syntaxique. Nous avons également utilisé un outil de programmation et d'environnement sémantique (*R*).

Ressources externes

Nous avons utilisé le thésaurus interne à la société ; il est organisé autour de six axes principaux : agroalimentaire, biens et services de consommation, industrie lourde, technologies de l'information et médias, sciences de la vie et services. Il couvre 450 secteurs d'activités économiques et est composé de 180 000 termes environ (le nombre exact fluctue assez régulièrement car le thésaurus est régulièrement mis à jour). Les traits sémantiques comme la POLYSEMY et la LINGUITIC-COMPLEXITY ont été déterminés par le thésaurus. La POLYSEMY correspond au nombre de fois qu'apparaît le terme dans le thésaurus (sous sa forme lemmatisée) et la LINGUITIC-COMPLEXITY correspond pour un terme donné au nombre de nœuds dans lequel il se place par rapport à l'arborescence du thésaurus.

Analyseur morpho-syntaxique

Nous avons utilisé l'analyseur morpho-syntaxique *xelda*, développé par *xerox*. Les grandes étapes de cet analyseur sont : il *(i)* identifie tout d'abord la langue (à partir des premiers caractères), *(ii)* segmente en phrases, *(iii)* tokénine (*i.e.* scinde un texte en unités lexicales élémentaires), *(iv)* analyse morphologiquement (renvoie les catégories grammaticales potentielles pour tous les mots identifiés durant la tokénisation) et enfin *(v)* désambiguïse morpho-syntaxiquement en déterminant la catégorie grammaticale d'un mot en fonction de son contexte.

Cet analyseur a été utilisé pour déterminer tous les traits morphologiques et syntaxiques. Exception pour déterminer le nombre de termes pour lequel un simple algorithme de calcul a alors été employé. Les *tokens* inconnus représentent les formes non recouvertes par *xelda* (mauvaise orthographe des termes par exemple). Les noms propres et les acronymes ne sont pas comptabilisés dans les *tokens* inconnus. Les étiquettes utilisées par *xelda* et utilisées pour cet analyse recouvrent principalement des noms (NOM), des adjectifs (ADJ), des noms propres (PROP). La liste entière des étiquettes est donnée en Annexe II (page II).

Le choix de l'analyseur se justifie par le fait qu'il est robuste et présente des résultats satisfaisants sur d'autres corpus. Cet analyseur est déjà utilisé et en place au sein de l'entreprise *Ubiquick* pour d'autres usages ; il présente notamment quelques erreurs d'étiquetage que nous avons pu régler lors de tests que nous avons rectifié. L'objectif ici n'est pas d'analyser les différences de comportement qu'il pourrait y avoir entre deux

analyseurs morpho-syntaxiques comme avec *Treetagger* [SCH 95] par exemple qui assez communément utilisé par la communauté scientifique en TAL), mais de bien de mettre en place une méthode de traitement des demandes en LN.

L'ensemble de notre chaîne d'analyse doit donc nous permettre de segmenter la demande en LN en plusieurs blocs d'informations afin d'en différencier les processus d'analyses. Le bloc [REFERENT] fera l'objet d'une analyse plus fine afin de pouvoir obtenir une représentation sémantique identique et donc comparable aux requêtes. Notre analyse présuppose l'étude de régularités lexicales d'une part, syntaxiques et énonciatives d'autre part, des deux types d'énoncés mais de façon différencié. C'est pour cela qu'elle s'appuie sur ces deux types d'analyses pour décrire les régularités dans les demandes en LN et dans les requêtes (occurrences des formes lexicales au sein de cadres syntaxiques et énonciatifs spécifiques, et à des places spécifiques à l'intérieur de ces cadres). Dans un deuxième temps, le but de cette analyse est de définir précisément si possible les caractéristiques qui permettent d'identifier spontanément un énoncé en fonction du profil utilisateur et de sa tâche de RI à effectuer : nous reconnaissons le discours de tel individu, de tel groupe, et nous pouvons donc appliquer des facettes pour sa recherche et essayer de mieux répondre à ses besoins informationnels. Ce sont les deux aspects que nous traiterons dans la section suivante.

Troisième partie

Résultats et Discussions

Notre expérimentation a pour objectif d'évaluer la façon dont un même utilisateur verbalise un besoin informationnel à travers une expression libre en LN puis un langage plus formalisé *i.e.* celui des requêtes avec mots-clés (ou équations booléennes pour les requêtes les plus élaborées). Nous en présentons les résultats dans cette 3ème partie (chapitres 6 et chapitres 7).

A priori, l'expression en LN est la plus longue dans le sens rédactionnel que les requêtes. N'étant pas limitée dans le nombre de caractères, l'expression en LN peut être exhaustive, permettant de contextualiser le besoin informationnel (thème de la recherche *i.e.* secteur d'activité, zone géographique, date, type de données, prix, etc.) sont autant d'éléments qui permettent de préciser la recherche). Nous expérimentons dans quelle mesure cette hypothèse est vraie. Par une expérience de terrain, nous étudions comment s'opèrent l'expression du besoin informationnel tout d'abord en LN puis à travers les requêtes. Comment sont exprimés dans ces deux types d'énoncés les éléments importants du besoin informationnel ? Sur quoi et comment s'opèrent le passage de la demande en LN aux requêtes ? Quelles sont les informations qui sont supprimées, ajoutées, modifiées entre les deux types d'énoncés ?

Cette expérience nous apporte des (premiers) éléments de réponse. Les résultats obtenus sont présentés selon deux principes dans les chapitres suivants :

Le chapitre 6 présente les corrélations entre des traits caractéristiques des demandes informationnelles en LN et les requêtes. Le chapitre 7 précise si les corrélations établies au chapitre 6 sont différenciées selon la tâche de RI à effectuer.

CHAPITRE 6

RÉSULTATS ET INTERPRÉTATIONS DES CARACTÉRISTIQUES DES DEMANDES EN LN ET DES REQUÊTES

Les résultats seront présentés selon trois axes :

1. les traits caractéristiques des demandes en LN afin de dégager d'éventuelles régularités comme la structure des demandes, les termes utilisés, les modes et temps grammaticaux employés, etc.

2. les traits caractéristiques des requêtes comme le nombre de requêtes par session utilisateur, le nombre de termes par requête ou encore les catégories grammaticales des termes, etc.

3. les points divergents et convergents entre les représentations sémantiques du besoin informationnel exprimé en LN et celles exprimé en langage de requêtes.

6.1 Traits caractéristiques des demandes en LN

Pour analyser les demandes en LN, nous avons croisé plusieurs analyses :

1. une analyse permettant la segmentation des demandes en LN en blocs d'informations ;

2. une analyse linguistique des informations contenues dans les différents blocs d'informations ;

3. une analyse morpho-syntaxique du « référent » des demandes en LN (voir à la page 70 pour une définition du « référent ») ;

4. et une analyse des demandes en LN par type de tâches de RI à effectuer.

6.1.1 Segmentation de la demande en LN en blocs d'informations

Nous avons pu remarquer à la section 5.1.1 que les demandes en LN s'articulent autour de plusieurs éléments, désignés comme autant de *blocs d'informations*. Comme nous l'avons déjà dit, ces blocs d'informations sont les suivants : [SALUTATION-DEBUT]

, [FONCTION-CLIENT], [INTENTION-RECHERCHE], [CONTEXTE], [TYPES-DONNÉES], [PRECISIONS], [REFERENT] et [SALUTATION-FIN]. Leur distribution au sein d'une demande en LN est présentée dans la Figure 6.1. Le nombre total de demandes en LN est de 1398.

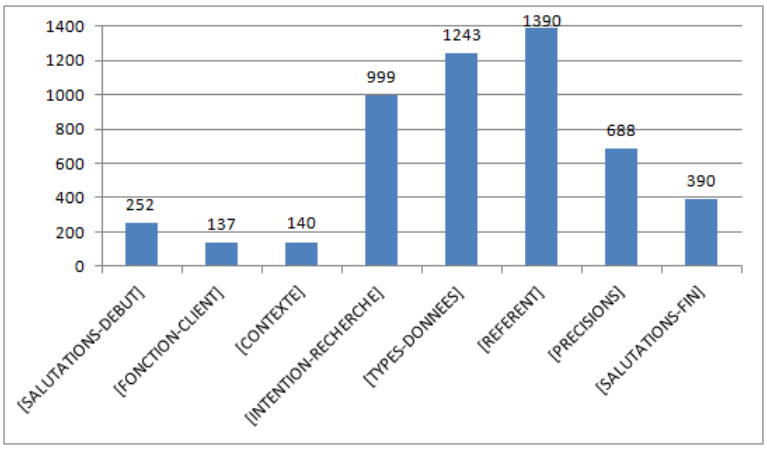

Figure 6.1 – Distribution des blocs d'informations sur les demandes en LN

Nous constatons que dans 1390 requêtes des demandes en LN (soit 99,42%) le concept [REFERENT] est renseigné. En effet, seulement quelques demandes (huit précisément) en LN ne mentionnent pas explicitement de référent : il s'agit de demandes beaucoup plus générales où les utilisateurs demandent des « secteurs d'activités qui marchent bien ». Les autres blocs sont bien renseignés : [TYPES-DONNÉES] à 88,91%, [INTENTION-RECHERCHE] à 71,45% ou encore le bloc [PRECISIONS] à 49,21%. D'autres concepts sont moins représentés comme le bloc sur la [FONCTION-CLIENT] à 9,79% ou encore ceux sur les [SALUTATIONS] (salutations-début à 18,02% ou salutations-fin à 27,89%). La distribution et ce découpage en blocs d'informations nous permet de mieux nous rendre compte de la structure et des éléments composants d'une demande en LN.

6.1.2 Traits linguistiques de la demande en LN

Nous présentons ci-dessous les principaux traits linguistiques des demandes en LN en détaillant (a) les traits morphologiques, (b) les traits syntaxiques et (c) les traits sémantiques. Les traits présentés sont inspirés des travaux de [MOT 05] sur la complexité d'une requête.

6.1.2.1 Traits morphologiques de la demande en LN

D'un point de vue morphologique, nous avons étudié plus particulièrement les points développés dans le Tableau 6.1 dont nous présentons les résultats ci-dessous. Nous présentons les résultats sous forme de moyennes (pourcentages) sur les 1398 demandes en LN.

Traits morphologiques de la demande en LN	
Longueur moyenne des demandes en LN	LENGTH
Moyenne des EN noms propres (pays)	PROPER_NAME_COUNTRY
Moyenne des EN noms propres (entreprises, marques)	PROPER_NAME_BRAND
Moyenne des acronymes (acro)	ACRONYM
Moyenne des valeurs numériques (dates)	NUM_DATE
Moyenne des valeurs numériques (quantités)	NUM_QUANTITY
Moyenne des valeurs numériques (prix)	NUM_PRICE

Tableau 6.1 – Traits morphologiques des demandes en LN inspirés de [MOT 05]

- **Longueur moyenne des demandes en LN [LENGHT]** : ce trait permet de relever la longueur des demandes. Les utilisateurs verbalisent leurs demandes avec 124 mots et en une seule phrase en moyenne. La demande la plus courte s'effectue en 6 mots tandis que la plus longue en 525 mots. L'écart type est assez important (80 mots) : il y a de grandes fluctuations dans la longueur des expressions des demandes en LN. Elles sont tantôt très explicites et développées avec de nombreuses informations de contextes, précisions, exemples notamment, tantôt être très succinctes dont la structure se rapproche alors à une requête avec seulement un ou plusieurs mots-clés sur le secteur d'activité recherché et une formulation de politesse. Exemples : « informations sur les biocapteurs » (29 caractères) ; ou encore « S'il

vous plait, j'aimerais avoir des études de marché sur Jeff de Bruges et merci ! » (69 caractères) ; « Bonjour, je suis un élève de l'ESC de Troyes qui prépare un exposé sur Nokia mais il me manque des données comme : la part de marches détenue par Nokia et ceux des ses concurrents, le chiffre d'affaire, les profits et bénéfices, le nombre d'employés Nokia. Merci par avance » (225 caractères).

- **Moyenne des noms propres (pays) utilisés [PROPER-NAME-COUNTRY]** : ce trait regroupe les noms propres des noms de pays (« France ») de villes ou de départements ainsi que leurs formes adjectivales (« français » dans « marché français » par exemple). Au total, 478 demandes (soit 34,19 %) contiennent au moins un lieu géographique. Cette information permet de déterminer une couverture géographique de la recherche.

- **Moyenne de noms propres : entités nommées des noms d'entreprises et / ou de marques [PROPER-NAME-BRAND]** : ce trait regroupe les noms de marques (« Coca Cola » ou d'entreprise (« Nestlé »). Les entités nommées des noms d'entreprises et ou de marques sont présentes dans 146 demandes (soit 10,44 %). Cette information permet de préciser sur quelle société l'utilisateur souhaite obtenir des données économiques (nombre de fusions ou acquisitions, bilans financiers, nombre de produits, etc), ou de citer une entreprise/marque comme exemple à partir de laquelle (desquelles) établir la recherche.

- **Moyenne des acronymes [ACRONYM]** : ce trait regroupe les acronymes présents dans les demandes en LN, comme U.E pour « Union Européenne ». Ils sont au nombre de 24 dans les demandes en LN soit 1,71 %. Ils sont donc peu présents mais peuvent représenter une certaine complexité lors de l'analyse (diversité du sens selon les contextes).

- **Moyenne des valeurs numériques : format dates [NUM-DATE]** : ce trait regroupe les mentions de temporalité comme l'année, ou le mois ainsi que la périodicité (« semestriel», « bimensuel »). Seulement 48 demandes soit 3,43 % mentionnent un critère de date dans la demande en LN. Les années de parution des études économiques (ex : publié en 2014) ou les couvertures temporelles (ex : prévisions sur 5 ans) ne semblent pas être un facteur important pour les utilisateurs de ce moteur de recherche.

- **Moyenne des valeurs numériques : format quantité [NUM-QAUNTITY]** : ce

trait est calculé en fonction de règles et de lexiques basés sur les nombreux cardinaux et certains symboles mathématiques comme le pourcentage. Dans l'expression « population 15 – 35 ans », la valeur « 15 – 35 » sera reconnue comme valeur numérique. Au total, 17 demandes soit 1,21% présentent des valeurs numériques de type quantité.

- **Moyenne de valeurs numériques (format prix) [NUM-PRICE]** : ce trait est calculé en fonction de règles et de lexiques permettant la reconnaissance d'entités numériques suivies ou précédées d'un signe monétaire. Dans l'expression « pas plus de 125 Euro », « 125 Euro » sera reconnu comme valeur numérique. Au total, 36 soit 2,56% des demandes en LN présentent un critère de prix, principalement lorsque l'utilisateur demande le tarif de la prestation.

6.1.2.2 Traits syntaxiques de la demande en LN

Les traits syntaxiques que nous avons étudiés sont basés à partir de [INTENTION-RECHERCHE] ; c'est en effet dans ce bloc d'informations que sont rassemblées les expressions d'intentions de recherche (« je voudrais faire un bilan sur », « je désire obtenir des informations sur »). Les traits syntaxiques des intentions de recherche via le bloc [INTENTION-RECHERCHE] de la demande en LN sont présentés dans le Tableau 6.2.

Traits syntaxiques de la demande en LN	
Moyenne des pronoms personnels	PP
Moyenne des structures syntaxiques correctes	STRUCT_SYNTAX_RIGHT
Moyenne des structures syntaxiques partielles ou incorrectes	STRUCT_SYNTAX_FALSE
Moyenne des phrases interrogatives	INTER_SENTENCE
Moyenne des phrases affirmatives ou négatives	AFFIR_NEG_SENTENCE

Tableau 6.2 – Traits Syntaxiques des demandes en LN inspirés de [MOT 05]

Parmi les traits syntaxiques, nous avons distingué :

- **l'usage des pronoms personnels [PP]** plusieurs pronoms personnels sont utilisés dans les demandes en LN : la première personne du singulier est utilisée dans 685 demandes soit 48,99%, la deuxième personne du pluriel dans 56 demandes soit 4,00%, la première personne du pluriel dans 52 demandes soit 3,71% et enfin la troisième personne du singulier dans 3 demandes soit 0,21% . La première personne du singulier est donc largement sur-représentée dans les demandes en LN représentant une démarche plutôt personnelle et individualisée.

- **les structures de phrases correctes [STUCT-SYNTAX-RIGHT]** : sur 1398 demandes en LN, 953 soit 68,16% ont une structure grammaticale correcte (comportant au moins un sujet, un verbe et un complément). Il s'agit de phrases déclaratives et interrogatives (ex : « Je veux des informations sur le marché du poisson surgelé à l'exportation depuis le Maroc »).

- **les structures de phrases incorrectes ou partielles [STRUCT-SYNTAX-FALSE]** : au regard des structures de phrases correctes qui étaient de 953 soit 68,16%, les 445 (31,83%) autres demandes sont soit *(a)* des « structures partielles » : ces structures partielles concernent 392 demandes soit 28,04% et sont principalement composées d'un verbe, d'un référent voire de quelques précisions mais pas de sujet. Exemple de structure partielle : « recherche étude sur le marché de la décoration exotique et orientale ». Soit *(b)* des « structures requêtes » : ces structures requêtes s'apparentent à des requêtes car seul un ou plusieurs mots-clés (identifiés dans le le bloc d'informations [REFERENT]) sont donnés. Ces structures requêtes concernent 53 demandes en LN soit 3,79% (ex : « marché épicerie fine » ou encore « étude de marché Champion »). Il est possible que l'expression des demandes en LN en structures-requêtes soit le fruit d' une erreur d'utilisation du formulaire SAV ; l'utilisateur peut penser encore utiliser le formulaire de recherche. Les structures requêtes concernent 3,79% des demandes en LN.

- **les phrases interrogatives [INTER-SENTENCE]** : parmi les demandes qui sont grammaticalement correctes (953 soit 68,16%), 270 (soit 19,31% sur l'ensemble des demandes en LN) sont interrogatives. Parmi ces phrases interrogatives, nous distinguons 186 (soit 13,30%) *interrogations totales i.e.* dont la réponse ne peut être que « oui » ou « non » (ex : « Auriez vous des documents sur les locations de voiture ? ») et 84 (soit 6,00%) *interrogations partielles i.e.* dont on ne peut pas

répondre que par « oui » ou « non » et qui porte sur un élément précis (ex : « Quelle est la part de marché de Nintendo ? »). Les interrogations partielles commencent par un mot interrogatif : adverbes ou pronoms interrogatifs[1]. Les *interrogations totales* sont donc deux fois plus nombreuses que les *interrogations partielles*.

- **les phrases affirmatives [AFFIR-SENTENCE] et négatives [NEG-SENTENCE]** : parmi les demandes qui sont grammaticalement correctes (953), une grande majorité sont affirmatives : 681 soit 71,45%. *A contrario*, seulement 7 phrases soit 0,73% des demandes sont négatives (*i.e.* contenant ne… pas, ne… plus, ne… jamais ou ne… point). Exemples : « je ne sais pas m'y prendre », « je ne trouve rien sur ce marché », « malheureusement notre fonctionnement ne nous permet pas de payer le prix affiché ». Ce n'est pas une forme couramment employée dans ce corpus qui pourtant relève d'un formulaire SAV ; les utilisateurs sont davantage dans une situation d'explication du besoin informationnel que de revendication.

D'autres traits, dont les variables connaissent une moins forte variance, peuvent également être à intéressants : sur l'ensemble des demandes en LN : l'indicatif est majoritairement utilisé : 654 demandes soit 46,78%, le conditionnel dans 270 demandes soit 19,31% et le gérondif dans seulement 6 demandes soit 0,43%. Par ailleurs, 153 demandes ont une structure impersonnelle (soit 10,94 %), 917 demandes sont au présent (soit 65,59%). Les autres temps sont quasiment inexistants : 9 demandes au passé (soit 0,64%), 4 au gérondif soit 0,29% et 2 au futur soit 0,14% ; le reste des phrases n'étant pas exprimé par un verbe conjugué.

Tous les verbes employés sont transitifs, 96 verbes sont pronominaux (soit 6,86%), 904 demandes (soit 64,66 %) ont des verbes à la voie active.

6.1.2.2.1 Résumé des principales tendances des demandes en LN Nous pouvons donc constater que certains traits caractéristiques de la demande en LN se détachent fortement comme la phrase déclarative affirmative, l'usage du présent de l'indicatif, la voie active ou encore l'utilisation de la première personne du singulier.

Parmi les variables qui connaissent la plus forte variance on peut relever : *(a)* les variables sur la longueur moyenne des demandes en LN, *(b)* la structure syntaxique de la demande, *(c)* la forme de la demande. Les variables qui connaissent les plus faibles

[1] Adverbes interrogatifs : « combien », « comment », « est-ce que », « n'est-ce pas », « pourquoi », « quand », « où », pronoms interrogatifs : « qui », « que », « quoi », « quel », « quelle », « quels », « quelles », « lequel », « lesquels »…

variances sont : *(a)* le temps du verbe, *(b)* le mode grammatical, et *(c)* le type de construction du verbe.

En conclusion, ce premier examen pour différencier au mieux les requêtes montre toutefois que les variances les plus fortes portent sur les variables qui sont directement impliquées dans l'énonciation (structure, forme et longueur des demandes en LN) ; il confirme la nécessité de repérer dans la demande en LN les éléments qui doivent fonctionner de façon contrastée sur ces variables.

6.1.2.3 Traits sémantiques de la demande en LN

Nous appelons ici lien sémantique tout rapport de signification pouvant exister entre deux termes distincts. Ainsi, dans le contexte informatique « langage de programmation » et « autocode » ont une relation sémantique. Les liens sémantiques entre deux termes peuvent être très variés et difficiles à exprimer de manière formalisée ; c'est pourquoi un certain nombre d'auteurs ont travaillé sur la désambiguïsation des termes en s'appuyant sur d'autres ressources comme WordNet et Wikipédia [SAN 10] [SAN 08], [LAL 11] : le principe est alors de rechercher les termes des requêtes dans divers dictionnaires afin de repérer les mots ayant plusieurs sens. Notre étude portant sur des secteurs économiques spécifiques, nous nous sommes appuyés sur les relations définies dans les thésaurus internes pour caractériser plus finement les liens sémantiques entre les termes.

Les traits sémantiques étudiés concernent : *(a)* la valeur polysémique et *(b)* la complexité linguistique des termes utilisés lors de la demande en LN dans le [REFERENT] et *(c)* l'usage des blocs CONTEXTE et PRECISIONS.

Ces traits sémantiques sont présentés dans le Tableau 6.3

- **la valeur polysémique** : cette valeur est calculée en fonction du nombre de fois que les termes apparaissent dans les thésaurus internes de la société. Plus un terme apparait dans les thésaurus, plus sa valeur polysémique est élevée (*i.e.* potentiellement mentionné dans différents secteurs d'activités donc non univoque). Ainsi, la *polysemy value* à 0 indique que le terme n'est mentionné qu'une seule fois dans les différents thésaurus ; il est donc peu polysémique. Les termes dont la *polysemy value* vaut 0 dans les référents sont majoritaires ; ils sont au nombre de 955 soit 68,31 %. Les *polysemy value* 1 à 3 sont présentées dans la Figure 6.2. Les termes dont la *polysemy value* vaut $k (k \geq 0)$ apparaissent $k+1$ fois sauf si $k = 3$ où le terme apparait au moins $k+1$ fois.

Traits sémantiques de la demande en LN	
Moyenne de la valeur polysémique (nbre de fois apparaissant dans le thésaurus) dans le bloc [REFERENT]	POLYSEMY
Moyenne de la complexité linguistique (profondeur des nœuds dans le thésaurus) dans le bloc [REFERENT]	LINGUISTIC_COMPLEXITY
Moyenne des usages des blocs [CONTEXTE] et [PRECISIONS]	CONTEXT_PRECISIONS

Tableau 6.3 – Traits sémantiques des demandes en LN inspirés de [MOT 05]

Les *[unknown]* sont les termes qui ne sont pas présents dans les thésaurus. Un *token* inconnu désigne une entité (ou unité) lexicale qui n'a pas été reconnue lors de l'analyse et les comparaisons avec les termes des thésaurus. Cette information peut soit indiquer que le terme a été mal orthographié faussant l'analyse avec la correspondance des termes issus des thésaurus, soit que le terme n'est tout simplement pas renseigné dans les thésaurus par manque de précision ou de recouvrement d'un secteur d'activité. Les token *[unknown]* sont au nombre de 105 soit 7,51%. Exemple : « marché du CMSystem » n'a pas été reconnu ; « CMSystem » peut soit renvoyer à une faute de frappe indiquant soit « Configuration management » soit « CRM system ». Ce terme a donc été étiqueté comme *[unknown]*. Autre exemple : « alliages de polymères » n'a pas été reconnu par manque de finesse dans le traitement : il renvoie en effet à un terme très spécifique en chimie qui n'a pas été traité dans les thésaurus (peut être ne recouvrant pas d'études économiques). Dans la figure 6.2, les Entités Nommées (EN), au nombre de 199, figurent indépendamment des token *[unknown]* ; en effet, les EN peuvent également être très polysémiques si elles ne sont pas contextualisées et doivent donc être comptabilisées dans les valeurs polysémiques. Exemple : selon que le terme soit un nom commun ou un nom propre « carrefour / Carrefour » (importance de la prise en compte de la majuscule), ce terme peut paraitre sans ambiguïté dans un contexte de supermarchés, grandes surfaces *versus* « je voudrais une étude au carrefour de plusieurs disciplines ». Les N|R pour non renseigné correspondent aux huit demandes sans référents.

- **la moyenne de la complexité linguistique** (profondeur des nœuds dans le thésaurus [LINGUISTIC-COMPLEXITY] : les [LEVEL 1] correspondent aux six axes principaux du thésaurus sectoriel : Agro-alimentaire, Biens et Services de Consommation, Industrie lourde, Technologies de l'information et Médias, Sciences de la Vie et Services. Les niveaux suivants de [LEVEL 2], [...] [LEVEL 5] sont des niveaux du thésaurus hiérarchiquement descendants. La distribution des [REFERENT] par level est illustrée dans la Figure 6.3 ; l'usage du thésaurus nous donne ainsi un aperçu de l'utilisation de la profondeur du nœuds. On voit notamment que ce sont les niveaux intermédiaires [LEVEL 3] et [LEVEL 4] sont les plus importants : 482 pour le [LEVEL 3] et 334 pour le [LEVEL 4] et dans une moindre mesure le [LEVEL 2] avec 185 demandes. Il y a peu de demandes concernant des secteurs généralistes (36 pour le [LEVEL 1]) ou des secteurs très spécifiques (95 pour le pour le [LEVEL 5]).

  Exemple de demande en LN dans le Tableau 6.4 : « Je suis un étudiant qui souhaite avoir des statistiques sur le marché de l'eau gazeuse. Merci d'avance » avec un référent de niveau [LEVEL 5] pour « eau gazeuse ».

- **la moyenne de l'usage des concepts** [CONTEXT-PRECISIONS]. En ce qui concerne le bloc [CONTEXTE], nous avons relevé trois emplacements possibles : [CONTEXTE-1] ; [CONTEXTE-2] ; [CONTEXTE-3] comme présenté dans la Figure 5.1 à la page 73. Dans 98 demandes en LN soit 7,01% un contexte est donné dans l'expression du besoin informationnel : 47 demandes le mentionnent dans le [CONTEXTE-1], 26 dans le [CONTEXTE-2] et 25 dans le [CONTEXTE-3]. Ces contextes permettent de spécifier le cadre du besoin informationnel : professionnel, personnel, universitaire ou autre. Citons par exemples : « Dans le cadre d'une implantation d'un restaurant », « dans le cadre d'une étude sur la qualité des approvisionnements et la traçabilité dans le secteur de la pâtisserie industrielle » ou encore « dans le cadre de ma formation ». En ce qui concerne le bloc [PRECISIONS] : 689 demandes en LN soit 49,28% contiennent des précisions. Il peut s'agir d'exemples « comme haricots blancs (Soissons) ; haricots rouges(marbré), pistache et pois du Cap », « ex : chez Exapaq » ou des détails sur le type de données désirées (« superficie », « chiffre d'affaire », « horaire d'ouverture », « nombre d'employés », « densité du réseau »).

Ces traits nous permettront ainsi de déterminer la difficulté sémantique de la demande en LN (par la complexité et la polysémie des termes). Cette difficulté de la de-

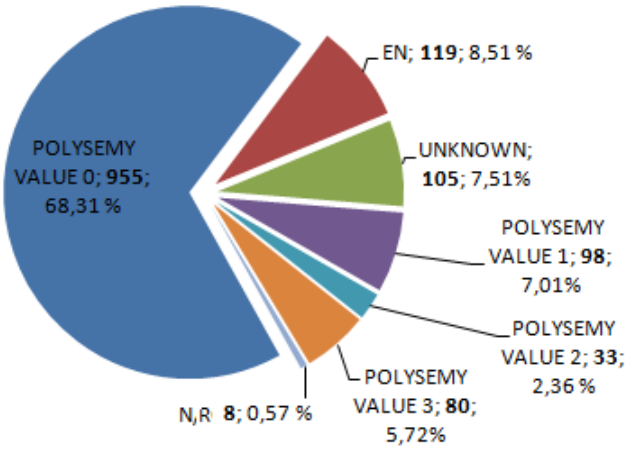

Figure 6.2 – Valeur polysémique des référents de la demande en LN

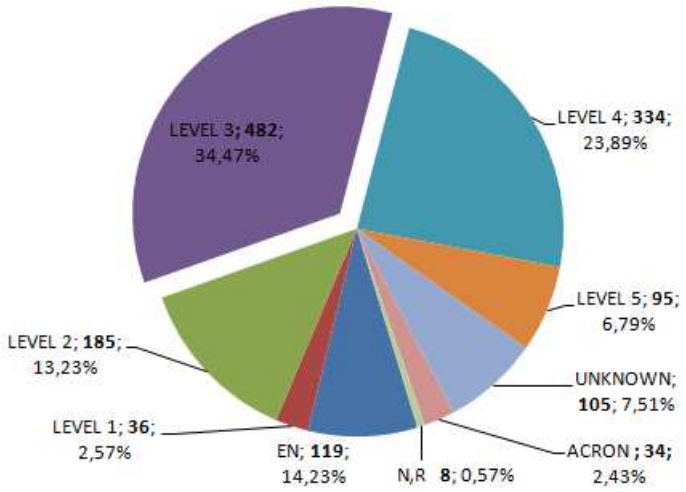

Figure 6.3 – Complexité linguistique des référents de la demande en LN et profondeur des nœuds dans le thésaurus

mande en LN peut également être évaluée par l'ajout ou non d'éléments contextuels. Cette contextualisation (c) se fait au travers des blocs CONTEXTE et PRECISIONS.

LEVEL 1		LEVEL 2		LEVEL 3		LEVEL 4		LEVEL 5	
2000	Agro-alimentaire	2013	Boissons	**2018**	**Boissons Non Alcolisées**				
2000	Agro-alimentaire	2013	Boissons	2018	Boissons Non Alcolisées	2019	Eau minérale	201983	Eau gazeuse
2000	Agro-alimentaire	2013	Boissons	2018	Boissons Non Alcolisées	2020	Boissons chaudes		
2000	Agro-alimentaire	2013	Boissons	2018	Boissons Non Alcolisées	2021	Jus		
2000	Agro-alimentaire	2013	Boissons	2018	Boissons Non Alcolisées	2022	Soda		
2000	Agro-alimentaire	2013	Boissons	2018	Boissons Non Alcolisées	2023	Boisson énergisante		

Tableau 6.4 – Extraction du thésaurus sectoriel pour le référent « eau gazeuse » de niveau [LEVEL 5].

6.1.3 Traits caractéristiques du référent dans les demandes en LN

A partir du découpage en blocs d'informations, nous avons effectué un travail plus spécifique sur les référents. En effet, comme nous l'avons vu à la sous-section 5.1.2.2 à la page 80, c'est dans ce bloc d'informations que se retrouve la plupart des éléments également formulées dans la requête du moteur de recherche. Le [REFERENT] est le concept porteur de l'information principale de la demande en LN ; il contient le thème de la recherche (secteur d'activité recherché). Afin de pouvoir être comparé plus finement avec la requête, ce bloc a fait l'objet d'une analyse qualitative avec une étude morpho-syntaxique de tous ses termes constitutifs. Pour cela, nous nous sommes basés sur les principaux traits relevés par [MOT 05] qui nous permettent d'avoir une vision plus précise sur la composition du [REFERENT]. Les traits linguistiques que nous avons repris et adaptés sont présentés dans le Tableau 6.5.

Nous avons étudié les variables suivantes :

- **Le nombre total de référents dans une demande en LN** : Ce référent peut lui même être découpé en plusieurs référents (lorsque l'utilisateur indique plusieurs thèmes) : [REFERENT-1], [REFERENT-2] ... allant jusqu'au [REFERENT-7].

  Exemple : [REFERENT-1] : « arts de la table » et [REFERENT-2] : « arts culinaires ».

- **Le nombre de termes par [REFERENT]** : on peut dénombrer 1390 référents issus de la demande en LN ; huit demandes en LN ne mentionnent pas de référents. Leur répartition est présentée dans la Figure 6.4 où nous pouvons remarquer que la majeure partie des utilisateurs s'expriment avec seulement 1 voire 2 référent(s). La longueur moyenne des référents est de 2,52 termes. La répartition du nombre de termes par référent est présentée dans la Figure 6.5. Il se détache de cette figure qu'une grande majorité des utilisateurs utilisent de 1 à 4 terme(s) par référent,

Traits syntaxiques du [REFERENT]	Type de n-grammes	Libellés
Profondeur syntaxique du bloc [REFERENT]		SYNT_DEPTH
Nom	Uni-grammes	NOUN
Adjectif	Uni-grammes	ADJ
Nom Propre	Uni-grammes	PROP
Nom + Nom	Bi-grammes	NOUN_NOUN
Nom + Adjectif	Bi-grammes	NOUN_ADJ
Nom Propre (x2)	Bi-grammes	PROP_PROP
Nom Préposition Nom	Bi-grammes	NOUN_PREP_NOUN
Nom Adjectif Nom	Bi-grammes	NOUN_ADJ_NOUN
Nom Propre (x3)	Tri-grammes	PROP_PROP_PROP
Nom (x 4)	Quadri-grammes	NOUN_NOUN_NOUN_NOUN
Nom Préposition Nom + 1	Quadri-grammes	NOUN_PREP_NOUN+1
Nom Propre (x4)	Quadri-grammes	PROP_PROP_PROP_PROP

Tableau 6.5 – Traits syntaxiques du bloc [REFERENT]

recouvrant ainsi 72% des demandes. A partir du référent 4, une cassure est assez nette sur le nombre de termes utilisé pour expliciter la demande en LN.

- **profondeur syntaxique du bloc [REFERENT] [SYNT-DEPTH]** : correspond au nombre total de n-grammes pour l'ensemble des référents (1 à 7). Les chiffres correspondant à cette information sont données dans la Figure 6.6. On voit ainsi apparaître que les uni-grammes sont majoritairement représentés dans les référents des demandes en LN. Les bi- et tri-grammes sont, dans une moindre mesure également assez présents dans le corpus. Ces trois types de n-grammes représentent ainsi 1563 référents sur les 1750 au total soit 89,31% des référents.

- **Analyse morpho-syntaxique des référents** : l'analyse morpho-syntaxique a été réalisée avec *Xelda* (voir la sous partie 5.3 pour une présentation de l'outil). Les résultats sont présentés dans les Figures 6.7 et 6.8. En ce qui concerne les **uni-grammes**, nous pouvons constater une nette dominance des noms (substantifs[2]) singuliers (étiquette NOUN sous *Xelda*) (57,46%) (ex : « tourime ») et dans une

[2]Unité lexicale (généralement un mot) qui désigne une chose ou une notion par elle-même.

moindre mesure des noms pluriel (20,02%) (ex : « déodorants »), totalisant un total de 623 référents soit 77,48%. Les noms propres (PROP) représentent une part non négligeable des référents (92 soit 11,44%) (ex : « Nokia »). Les adjectifs (ADJ) (singuliers et pluriels) représentent 55 référents soit 6,84% (ex : « bancaire »).

En ce qui concerne les **bi-grammes** : ceux qui contiennent un nom sont les plus représentés. Ils découlent des uni-grammes soit par l'ajout d'un adjectif qui le suit [NOUN ADJ] (306 soit 70,02%) (ex : « pierre naturelle ») ou qui le précède [ADJ NOUN] (11 soit 2,52%) (ex : « petit commerce »), soit d'un nom commun (72 soit 16,48%) (ex : « rayon fruit ») ou d'un nom propre (14 soit 3,20%) (ex : « Sunny Delight »).

En ce qui concerne les **tri-grammes** : ils sont principalement représentés par des « nom [de] nom », présents dans 250 référents soit 77,63% des tri-grammes (ex : « espace de beauté », « vernissage de bois », « transport de colis », « fabrication de pain »). Les noms suivis d'un adjectif [NOUN ADJ ADJ] ou [NOUN ADJ NOUN] représentent quant à eux 26 référents soit 8,07% (ex : « produits diététiques naturels » ou encore « produits alimentaires gourmet »). Les autres formes sont beaucoup moins fréquentes.

En ce qui concerne les **quadri-grammes** : tout comme dans les tri-grammes, les n-grammes les plus représentés sont les « nom [de] nom » avec l'ajout d'un déterminant ou d'un adjectif (« N1 Prep det N2 »,« N1 [de] det N2 »,« N1 [de] [le] N2 », « N1 ADJ DET N2» pour n'en citer que quelques-uns[3] (ex : « parc automobile des entreprises », « recyclage de matériel informatique », « jus de pomme artisanal »). Cette forme représente 72 référents des quadri-grammes soit 68,57% d'entre eux. Sur le même modèle des tri-grammes, les quadri-grammes basés sur la forme [NOUN ADJ NOUN] représentent 20 référents soit 19,04% des quadri-grammes.

Nous n'avons pas analysé les syntagmes composés de 5 termes et plus car il était difficile de d'en extraire des caractéristiques communes et modélisables.

[3]Nous avons travaillé sur des « filtres souples » pour collecter les variantes de la dénomination « nom [de] nom » que nous avons considéré comme constitutif d'une « familles d'arbres ». Ces « familles d'arbres » sont en effet construits sur le même modèle et il aurait donc eu une extrême redondance à les énumérer. Nous avons préférer restituer les principaux n-grammes notamment ceux émanant des structures déjà identifiés lors de l'analyse des tri-grammes.

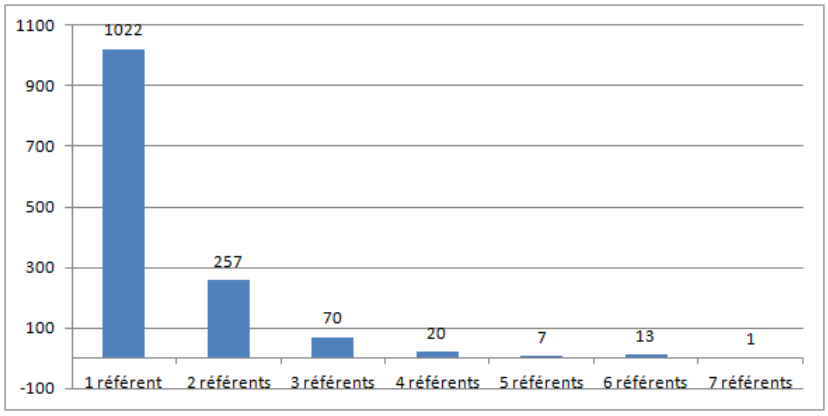

Figure 6.4 – Nombre de concepts [REFERENT] dans les demandes en LN

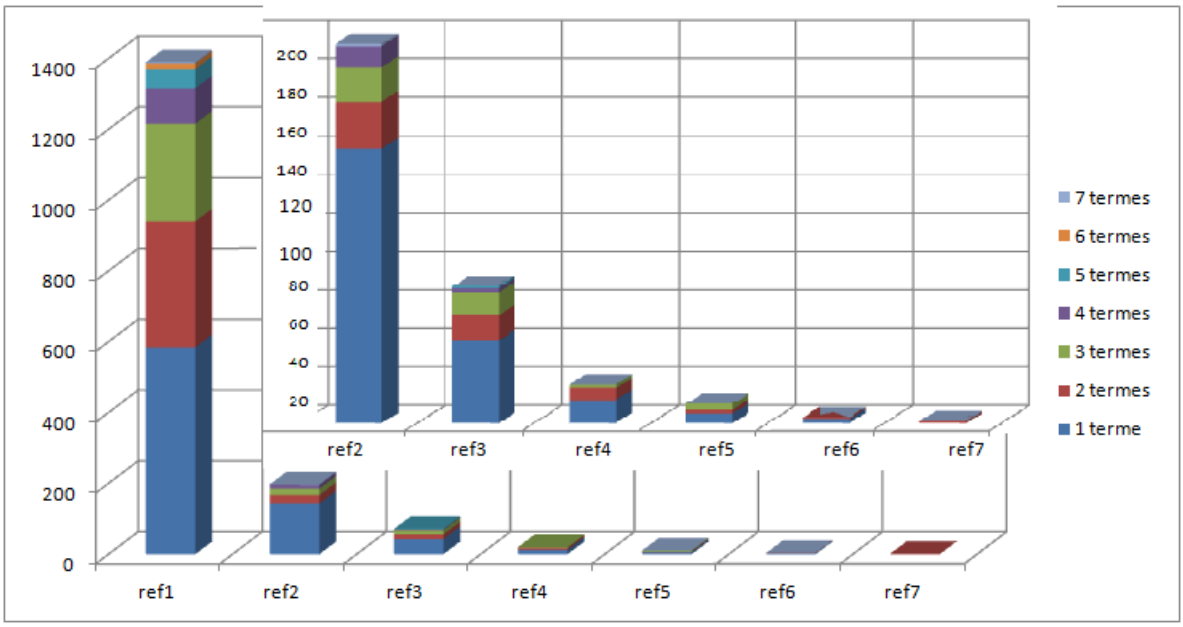

Figure 6.5 – Nombre de termes par [REFERENTS] dans les demandes en LN

6.2 Traits caractéristiques des requêtes

Nous étudions ici les traits caractéristiques des requêtes à partir des *logs* de requête(s). Pour rappel, dans le cadre de la recherche d'informations sur le Web, les journaux ou *logs* de requête(s) *« search logs »* correspondent aux fichiers de type journal et au format texte contenant les échanges communicatifs entre un système et ses utili-

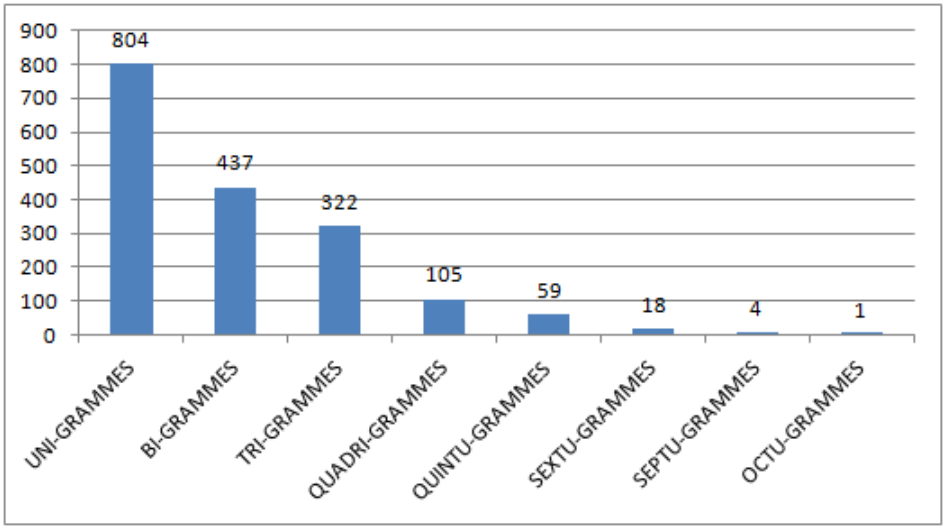

Figure 6.6 – Nombre de n-grammes par référents dans les demandes en LN

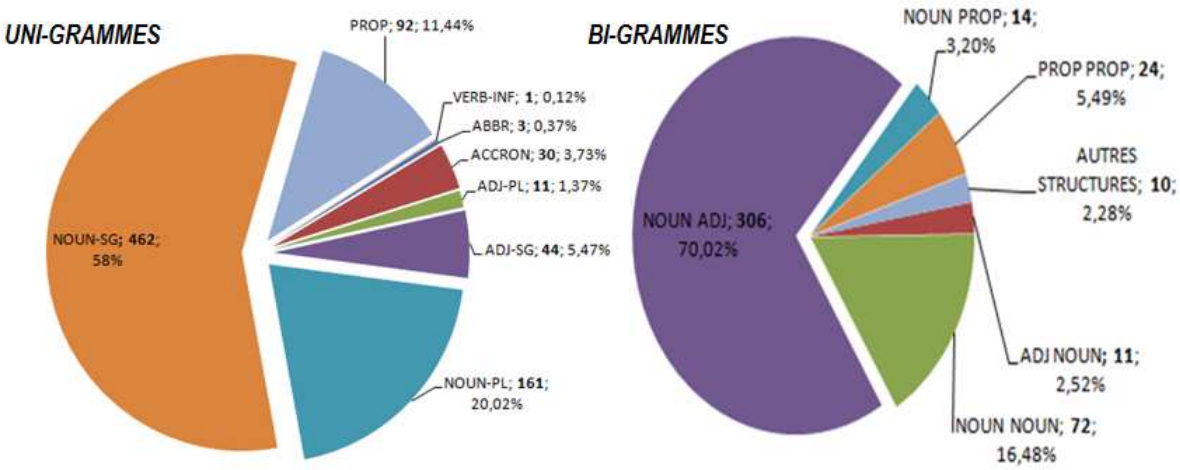

Figure 6.7 – Étiquetage morpho-syntaxique des uni- et bi-grammes dans les [REFERENTS] des demandes en LN

sateurs. Ces fichiers enregistrent un certain nombre d'informations comme l'adresse IP, le fichier atteint, la date et l'heure de connexion, les requêtes saisies dans le moteur de recherche, la date et l'heure de ces requêtes ainsi que le nombre de résultats obtenus. Parmi ces informations, nous exploitons certaines statistiques comme *(a)* le nombre de requêtes par session utilisateur, *(b)* la longueur des requêtes, *(c)* ; des analyses morpho-syntaxiques des termes et enfin *(d)* des spécificités des équations de recherche (présence

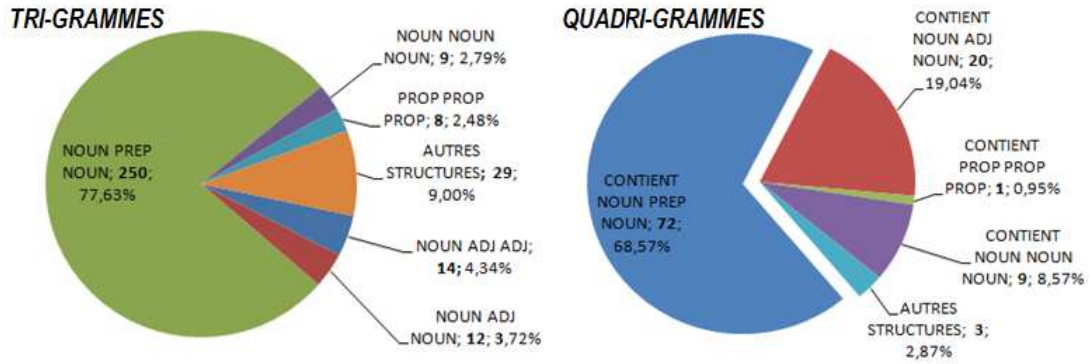

Figure 6.8 – Étiquetage morpho-syntaxique des tri- et quadri-grammes dans les [REFERENTS] des demandes en LN

d'opérateurs booléens, d'entités nommées, de dates, de lieux géographiques, etc).

6.2.1 Statistiques sur les requêtes

L'utilisation du moteur de recherche `Plusdetudes.com` se fait par l'intermédiaire d'une inscription obligatoire. L'ouverture d'une session est explicite avec la saisie d'un nom et d'un mot de passe. Nous avons considéré dans une session utilisateur la demande en LN suivie d'une ou plusieurs requête(s). Parmi cette session, nous avons relevé :

- le nombre de requêtes par session utilisateur,
- la longueur moyenne des requêtes par session utilisateur,

Nombre de requêtes par session utilisateur

La répartition des 1398 demandes en LN en nombres de requêtes par session utilisateur est présentée dans la Figure 6.9. On peut ainsi constater que l'expression du besoin informationnel s'exprime le plus souvent en une requête (42,8%) voire deux requêtes (24,89%). *A contrario*, il ne s'exprime que très rarement avec 5 requêtes ou plus (155 requêtes soit 11,08 %). D'après la Figure 6.9, on peut ainsi constater qu'un utilisateur a formulé 33 requêtes pour un même besoin informationnel (40 dans [CHA 05]). Ce nombre de requêtes reste tout de même extrêmement rare.

Le nombre total de requêtes est de 3422 requêtes pour les 1398 utilisateurs. Le résultat que nous avons obtenu est également cohérent avec ceux d'autres études comme [SIL 99] et [SPI 01].

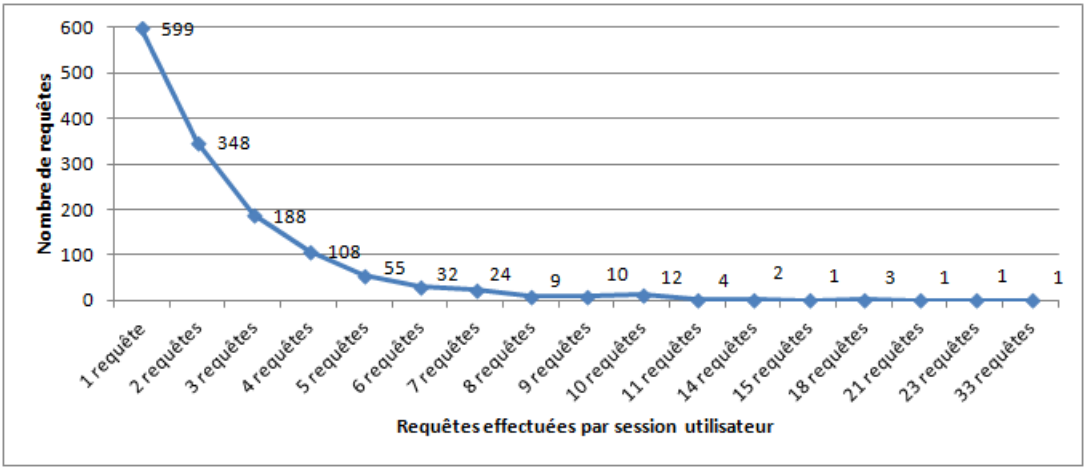

Figure 6.9 – Répartition des 1398 demandes en LN en nombre de requêtes par session utilisateur

Longueur moyenne des requêtes

La longueur moyenne des requêtes (nombre de termes) est de 2,4 termes par requête. Cette moyenne est représentée dans la Figure 6.10 pour les 3422 requêtes. Ce résultat corrobore les travaux de [DEL 05] par exemple sur un moteur Intranet (Arianet) avec 2,13 termes en moyenne par requête, ceux de [SIL 99] déjà anciens sur le moteur de recherche Altavista avec une moyenne de 2,35, ceux de [SPI 01] sur le moteur Excite avec une moyenne de 2,21 ou encore 2,8 à partir des travaux de [SPI 02] ou encore ceux de [CHA 05] d'après les *logs* de requête du moteur de recherche du gouvernement de l'Utah.

On remarque qu'au-delà de la quinzième requête effectuée par l'utilisateur, le nombre de termes par requête est beaucoup plus fluctuant car basé sur un trop petit échantillon pour être correctement analysé.

Nombre de n-grammes dans les requêtes

Pour cela, nous avons analysé toutes les requêtes composées de 1 à 4 termes soit 1572 uni-grammes, 831 bi-grammes, 606 tri-grammes et 296 quadri-grammes.

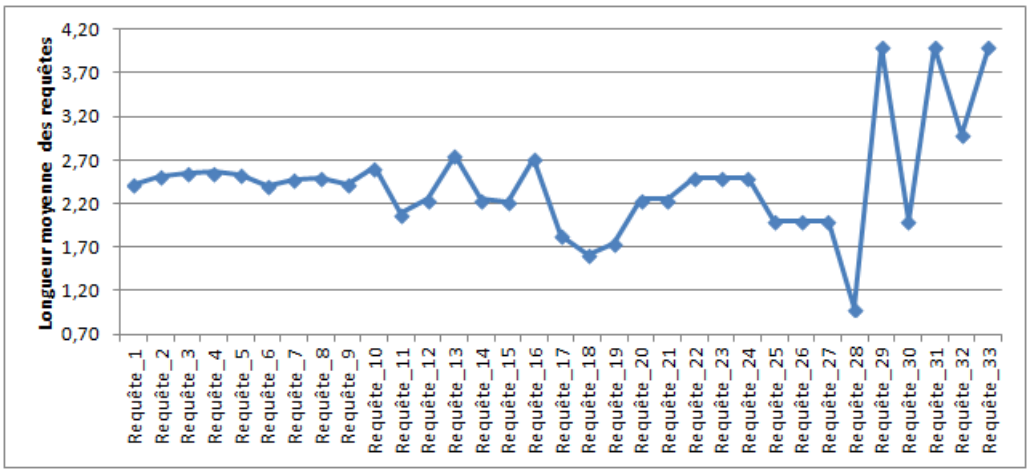

Figure 6.10 – Longueur des requêtes

6.2.2 Catégories morpho-syntaxique des requêtes

Nous nous intéressons ici au degré d'explicitation de l'intention des utilisateurs à travers leurs requêtes. Pour cela, nous avons analysé toutes les requêtes composées de 1 à 4 termes soit 1572 uni-grammes, 831 bi-grammes, 606 tri-grammes et 296 quadri-grammes.

Cette analyse s'est effectuée avec l'analyseur morpho-syntaxique *Xelda* pour déterminer leur appartenance aux catégories morpho-syntaxiques. Les résultats de cette analyse sont présentés dans les Figures 6.11 et 6.12. Au delà (5 termes et plus), les structures étaient trop peu nombreuses pour avoir des chiffres intéressants à analyser et nous ne les avons pas retenues pour l'analyse.

Nous remarquons que dans les catégories morpho-syntaxiques les plus utilisées pour formuler les requêtes sont les noms (NOUN) et les noms propres (PROP)[4].

6.2.2.0.1 Les requêtes uni-grammes

- **les NOUN dans les requêtes uni-grammes** : le NOUN représente 1196 des requêtes uni-grammes sur les 1572 requêtes au total soit 76,08%. Exemples : « assurance », « décoration », « voitures », « cosmétiques ».

- **les PROP (proper name) dans les requêtes uni-grammes** : 219 requêtes uni-grammes sont des PROP soit 13,93%. L'analyseur distingue quatre types de PROP :

[4]Il s'agit ici des libellés *Xelda*.

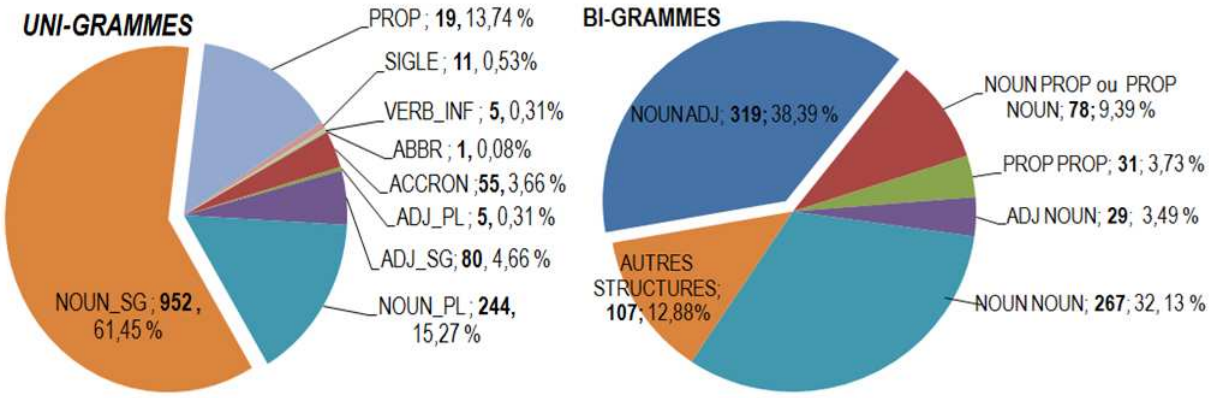

Figure 6.11 – Étiquetage morpho-syntaxique des uni- et bi-grammes dans les requêtes

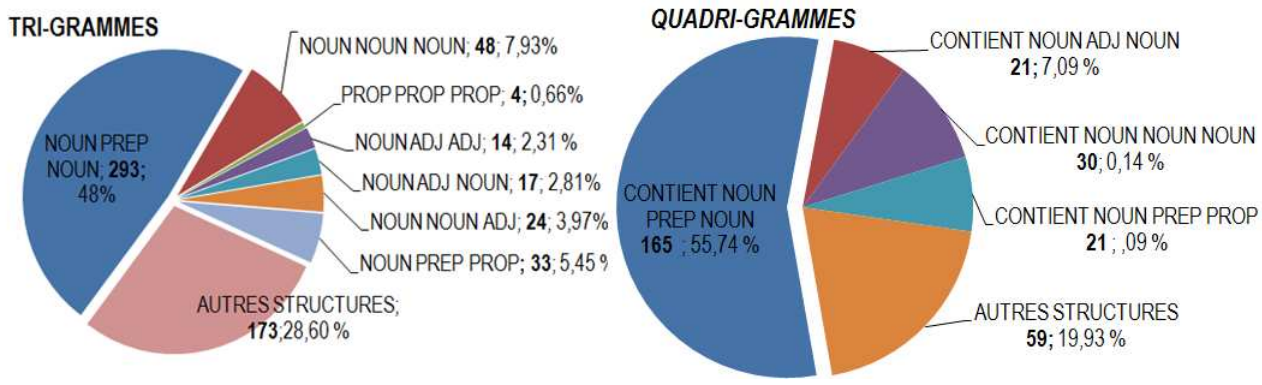

Figure 6.12 – Étiquetage morpho-syntaxique des tri- et quadri-grammes dans les requêtes

(a) les noms propres représentant des noms de marques ou sociétés (*prop +Bus*)(business name)(ex : « Amazon » ou « Nokia »), *(b)* les noms propres représentant les noms de continents (*prop + Continent*)(continent name) (ex :« Europe »), *(c)* les noms propres représentant des noms de pays, régions ou états (*prop +Country)* (country name) (ex : « Espagne ») et enfin *(d)* les noms propres représentant plus spécifiquement les noms de villes (*prop +City*) (city name) (ex : «Montpellier»).

- **les ADJ dans les requêtes uni-grammes** : les adjectifs sont aussi présents dans les requêtes uni-grammes : 85 (singulier ou pluriel) soit 5,40% des requêtes uni-grammes (ex : « interactif »).

6.2.2.0.2 Les requêtes bi-grammes

- **les NOUN dans les requêtes bi-grammes** : le NOUN apparaît soit accompagné d'un adjectif (« NOUN ADJ » ou « ADJ NOUN »), c'est le cas dans 348 des requêtes bi-grammes sur les 831 (soit 41,87%) au total avec par exemple : « emballages souples ». Le NOUN peut également être accompagné d'un autre nom dans les requêtes bi-grammes : « NOUN NOUN » ; c'est le cas dans 267 (soit 32,12%) requêtes bi-grammes avec par exemple « aliments santé » ou « alarmes intrusion ». Et enfin, le NOUN peut être accompagné d'un nom propre : « NOUN PROP » ou « PROP NOUN» : c'est le cas dans 78 (soit 9,38%)requêtes bi-grammes comme avec par exemple « sécurité Sensormatic »).

- **les PROP dans les requêtes bi-grammes** : les PROP dans les requêtes bi-grammes sont présentes dans les structures suivantes : « NOUN PROP », « PROP NOUN» ou « PROP PROP » dans 29 requêtes soit 3,48%. Il peut s'agir de combinaisons des différentes PROP comme *+Bus + City* (ex :« Carrefour Montpellier»), *+Bus +Continent* (« Amazon Europe ») ou encore *+Bus +Bus* (ex : « Segafredo Zanetti »)

6.2.2.0.3 Les requêtes tri-grammes

- **les NOUN dans les requêtes tri-grammes** : sont présents dans les structures compléments du nom telles que NOUN PREP NOUN qui englobent les « nom [de] nom » (ex : « logiciels de gestion ») ou encore les « nom [à] nom » (ex : « stylo à bille »). Ces structures NOUN PREP NOUN sont au nombre de 293 requêtes soit 48,34% de l'ensemble des requêtes tri-grammes (606 au total). Les NOUN sont également présents dans d'autres structures tri-grammes telles que : NOUN NOUN NOUN avec 48 requêtes tri-grammes soit 7,92% (ex : « société nettoyage entretien ») ; NOUN ADJ NOUN avec 17 requêtes tri-grammes soit 2,80% (ex : « librairie spécialisée jeunesse ») ; NOUN ADJ ADJ avec 14 requêtes tri-grammes soit 2,31% (ex : « produits alimentaires régionaux ») et enfin NOUN NOUN ADJ avec 24 requêtes tri-grammes soit 3,96% (ex : « maintenance réparation informatique »).

- **les PROP dans les requêtes tri-grammes** : les PROP dans les requêtes tri-grammes interviennent dans deux structures : les NOUN PREP PROP ou les PROP PROP PROP. Les NOUN PREP PROP sont au nombre de 33 soit 5,44% des requêtes tri-grammes (ex : « restauration à Paris » ou « cosmétique en Inde »). Les PROP dans la structure

NOUN PREP PROP peuvent être des *+Bus, +City, +Continent*. Les trois catégories sont regroupées sous PROP. Les structures PROP PROP PROP sont beaucoup moins représentatives puisqu'elles sont au nombre de 4 (0,66%) (ex : « Dell Hewlett Packard » ou encore « Nestlé After Eight »). Il peut s'agir là encore, comme pour les exemples présentés d'une combinaison ou de duplications d'une des catégories PROP suivantes : *+Bus +City +Continent*.

6.2.2.0.4 Les requêtes quadri-grammes

- **les NOUN dans les requêtes quadri-grammes** : apparaissent également dans les structures telles que NOUN PREP NOUN +1 où « +1 » signifie soit *(a)* qu'un nom peut être placé avant le NOUN PREP NOUN comme dans « matériel planche à voile » ou après comme dans « supports de communication tarifs » soit *(b)* qu'il s'agit d'un ajout d'un article (ex : « marché de la décoration ») soit *(c)* qu'il s'agit d'un ajout d'un adjectif (ex : « vente de matériel informatique »). Il était intéressant pour notre étude de comptabiliser les structures issues de NOUN PREP NOUN afin de pouvoir les corréler avec les requêtes tri-grammes de ce type. Les NOUN PREP NOUN +1 sont au nombre de 165 dans les requêtes quadri-grammes (soit 55,74% sur l'ensemble des 296 requêtes quadri-grammes). Le NOUN apparait dans les requêtes quadri-grammes dans les structures telles que NOUN ADJ NOUN +1 (ex : « marché électronique ingénierie informatique ») dans 21 requêtes quadri-grammes (soit 7,09%) ou encore dans NOUN PREP PROP +1 (ex : « secteur quincaillerie au Mexique » dans également 21 requêtes quadri-grammes (soit 7,09%). Les NOUN NOUN NOUN +1 sont présents dans 30 requêtes quadri-grammes (10,14%) (ex : « vidange pompage entretien fosses »).

- **les PROP dans les requêtes quadri-grammes** : peuvent être présentes dans les NOUN PREP PROP +1, NOUN PREP NOUN +1. Très peu d'autres structures ont été relevées avec les PROP dans les requêtes quadri-grammes et sont très disparates ; elles sont alors comptabilisées dans 'autres structures' comme illustré dans la Figure 6.12.

En conclusion, le NOUN apparaît dans 1196 sur 1572 requêtes uni-grammes (soit 76,08 %), dans 693 sur 831 requêtes bi-grammes (soit 83,39 %), dans 415 sur 606 requêtes tri-grammes (soit 68,88 %) et enfin dans 237 sur 296 requêtes quadri-grammes (soit 80,06 %). Il apparait comme la catégorie grammaticale la plus utilisée dans les

requêtes comme indiqué notamment par [LI 10], suivi par les PROP et les formes adjectivales accompagnées d'un nom.

Dans leur étude, [BAR 08] ont analysé les requêtes du moteur de recherche *Yahoo !* en Aout 2006. Sur 2 508 requêtes uni-grammes, 40,2% sont des noms propres et 30,9% sont des noms, 7,1 % des adjectifs. La distribution de ces catégories morpho-syntaxique est un peu différente car la part des noms propres est un peu plus importante.

6.2.3 Spécificités des équations de recherche du moteur

Nous nous sommes intéressés aux traits caractéristiques et spécifiques des requêtes pour plusieurs raisons :

1. La première est de savoir comment les utilisateurs de `Plusdetudes.com` perçoivent et utilisent le moteur de recherche : utilisation ou non d'opérateurs booléens (ET, OU, SAUF), des guillemets, des signes diacritiques (majuscules, accents) ou encore le recours à une langue étrangère pour formuler leur besoin ;

2. La seconde est d'appréhender la façon dont les utilisateurs formulent et explicitent le secteur d'activité recherché : Contient-elle des indications géographiques ? temporelles ? Des indications spécifiques comme des entités nommées (noms d'entreprises, d'organisations ou encore de personnes) ou encore le type de données recherchées ?

Nous abordons ces différents aspects dans les paragraphes suivants.

L'usage des opérateurs booléens dans les requêtes [BOOLEAN-OPERATOR] Nous avons ici relevé les opérateurs booléens (ET, OU, SAUF), les opérateurs de troncature (termes de la requête finissant par une astérisque '*'), et les guillemets utilisés dans les requêtes. Sur la totalité des requêtes (soit 3422), 382 d'entre elles contenaient au moins un opérateur booléen. Le plus représenté de ces opérateurs est le ET (représenté aussi par un '+' et traité de la même façon dans le moteur de recherche `Plusdetudes.com`) pour les deux tiers d'entre eux. Les guillemets, les troncatures et l'opérateur booléen SAUF sont plus rarement utilisés. L'opérateur OU n'a jamais été utilisé dans notre corpus. La distribution de l'utilisation des opérateurs booléens est représentée dans le Tableau 6.6.

A partir de la 10ème requête, il est difficile de tirer des conclusions sur l'usage des opérateurs booléens car les données sont trop réduites (moins de 32 requêtes). Néan-

Num. requête	Nbre total de Requêtes	Requêtes avec [BOOLEAN-OPERATOR]	%
Requête 1	1398	51	3,64%
Requête 2	789	101	12,65%
Requête 3	452	74	16,37%
Requête 4	263	43	16,34%
Requête 5	156	32	20,51%
Requête 6	101	18	17,82%
Requête 7	68	13	19,11%
Requête 8	44	7	15,90%
Requête 9	34	5	14,70%

Tableau 6.6 – Utilisation des opérateurs booléens dans les requêtes 1 à 9

moins, on peut constater, d'après la Figure 6.13 que le ratio entre l'usage des opérateurs booléens et les requêtes est très important en cas de multi-requêtage *i.e* plusieurs recherches effectuées sur le moteur de recherche pour un même utilisateur et une même session de recherche. On constate d'après ce ratio, trois phases de comportements de recherche : *(a)* la première requête (la seule chez 599 utilisateurs soit 42,84%) s'effectue peu avec des opérateurs booléens, *(b)* les requêtes suivantes (requête 2 à requête 9) indiquent plus généralement l'introduction d'un opérateur booléen et *(c)* la troisième phase en cas de nombreux multi-requêtages (à partir de la requête 10) montre un usage assez important des opérateurs booléens. Les résultats de cette troisième phase ne conduisent pas à une conclusion très robuste : ils représentent en effet seulement 86 requêtes. Sur ces 86 requêtes, 32 contenaient un opérateur booléen ce qui est assez notable.

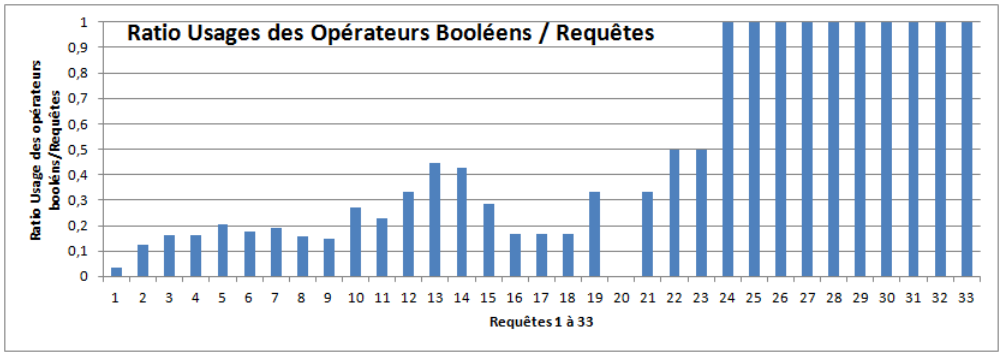

Figure 6.13 – Ratio de l'usage des opérateurs booléens en fonction du nombre de requêtes

L'utilisation de signes diacritiques (les majuscules et accents) L'utilisation des signes diacritiques (comme les accents) peuvent modifier considérablement les résultats de la recherche ; ce qui est particulièrement vrai pour les accents dans ce moteur. Or, environ 30% des requêtes qui avaient besoin d'accent en souffraient et un travail orthographique important a du être mis en œuvre afin de pouvoir exploiter les résultats. L'usage des parenthèses dans le moteur de recherche est par contre indifférencié (il n'est pas *case sensitive*) mais 9% des requêtes étaient constituées de majuscules. Ces chiffres indiquent que les utilisateurs ont une certaine perception de ces signes diacritiques qu'ils interprètent comme une règle d'usage pour la recherche d'informations à travers un SRI.

L'utilisation d'une langue étrangère dans la requête [FOREIGN-LANGAGE] Très peu de requêtes sont formulées en langue étrangère : 108 requêtes sur les 3422 soit 3,15% (Figure 6.14). Toutes sont en anglais. Ce chiffre est relativement faible mais notons que ce moteur est exclusivement présenté en français tout comme les documents renvoyés à l'utilisateur.

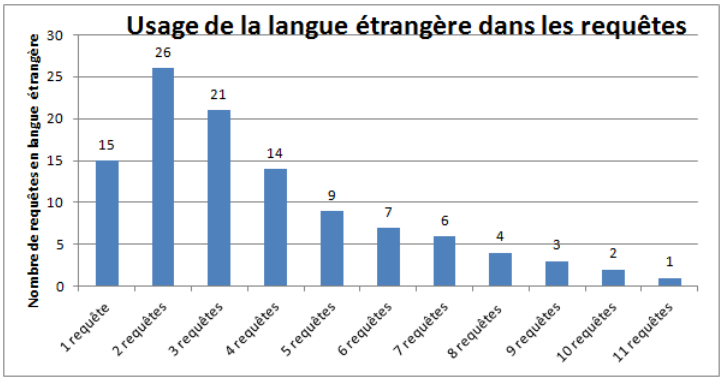

Figure 6.14 – Usage de la langue étrangère dans les requêtes

Tout comme les opérateurs booléens, le recours à une langue étrangère est plus importante dans les cas de multi-requêtages plutôt que dans les cas d'une première requête (Figure 6.15).

Les utilisateurs modifient leurs stratégies de recherche et formulent leurs requêtes en langue étrangère souvent après leur première phase de recherche. On peut supposer que les résultats obtenus lors de cette première phase ne répondaient que partiellement aux attentes des utilisateurs.

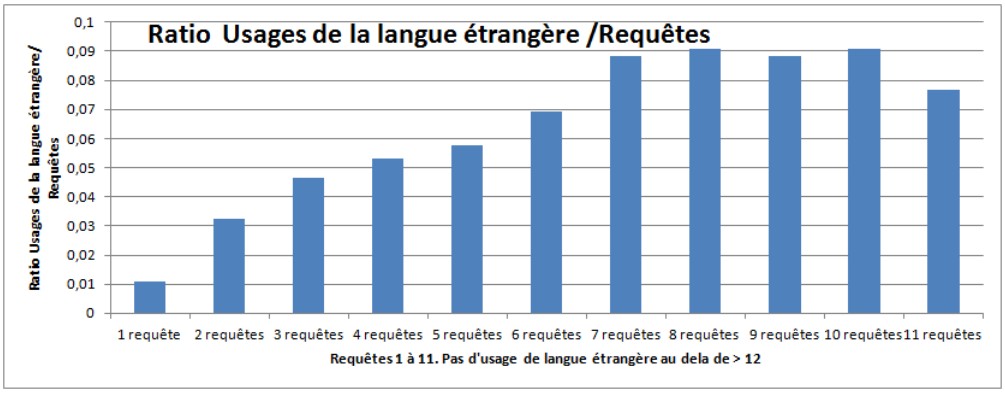

Figure 6.15 – Ratio de l'usage de la langue étrangère en fonction du nombre de requêtes

Les indications temporelles dans la requête [NUM-DATE] La date désiré pour une étude lors de la phase de recherche n'est que très rarement mentionnée : sur les 3422 requêtes totales, seulement 5 dates sont indiquées dans les quatre premières requêtes (ex : « marché du champagne en 2002/2003 », « ameublement 2005 »). Au delà de ces quatre premières requêtes, il n'y a plus aucune indication temporelle dans les requêtes. C'est un résultat plutôt étonnant car la date est un élément important dans les études économiques. Nous pouvons imaginer que les utilisateurs s'attendent à n'avoir en résultats que des études récentes et qu'il n'est donc pas nécessaire d'indiquer une date spécifique pour avoir les résultats les plus récents.

Les indications géographiques dans la requête [PROPER-NAME-COUNTRY] La zone géographique est indiquée dans 393 requêtes sur les 3422 requêtes au total soit 11,48%. Il peut s'agir soit de noms de villes (principalement en France), identifiées par les étiquettes *Xelda* par *+City* (ex :« fréquentation des restaurants sur Bordeaux »), soit des noms de pays *+Country* (ex :« décoration Royaume-Uni »), soit des noms de continents +Continent (ex : « vin en Europe ». Les requêtes mentionnant exclusivement un nom de pays, de continent ou de ville sont également comptabilisés (ex : « Asie »). Cette indication est plus importante dans des cas de multi-requêtages à partir de la 2ème requête que lors de la première requête. Dans le Tableau 6.7, on peut constater que lors de la première requête, la géographie est mentionnée dans 8,51% des cas alors qu'elle passe entre 12 à 22% lors de multi-requêtages. Cette hausse est probablement dûe aux résultats retournés par le moteur : en effet, ce dernier donne des études de marché globales, c'est

à dire couvrant tous les pays du monde lorsqu'aucun pays n'est explicitement mentionné dans le formulaire de recherche.

Num. requête	Nbre total de Requêtes	Requêtes avec [NUM-DATE]	%
Requête 1	1398	119	8,51%
Requête 2	789	98	12,42%
Requête 3	452	62	13,71%
Requête 4	263	42	15,96%
Requête 5	156	21	13,46%
Requête 6	101	14	13,86%
Requête 7	68	11	16,17%
Requête 8	44	8	18,18%
Requête 9	34	4	14,70%
Requête 10	22	5	22,72%
+ de 10 requêtes	86	9	10,46%

Tableau 6.7 – Indication de la zone géographique dans les requêtes

Les indications sur le type de données recherchées dans la requête [DATA-TYPE] Le type d'informations recherchées est indiqué dans 651 requêtes sur 3422 requêtes au total soit 19,02%. Parmi les types de données les plus recherchés, nous relevons : « étude(s) de marché », « étude(s) de faisabilité », « base(s) de données », « chiffre d'affaires » / « CA », « force de vente », « forces et faiblesses », « statistiques du marché », « évolution du marché », « analyse de bilan », « forces et faiblesses », « études qualitatives », « fusions et acquisitions », « offre(s) et demande(s) », « prévision(s) de vente »,« étude marketing »... Quasiment une requête sur cinq contient donc ce type d'informations. Cette donnée est également très présente dans les cas de multi-requêtages comme le montre les résultats du Tableau 6.8.

Les indications de type Entités Nommées (Business) dans la requête [PROPER-NAME-BUS] Le nombre de requêtes contenant des entités nommées de type *+Bus* (business name) est de 287 sur les 3422 requêtes au total soit 8,38 %. Ces entités nommées concernent des noms d'entreprises (« Nestlé », « Cora », « Pernod Ricard ») mais aucun de noms de personnes de type « +Farm » (last name). Les entreprises les plus récurrentes dans le corpus sont l'« Oréal », « Coca-Cola », « Nokia », « Danone », « Carrefour », « Amazon » ; grandes marques de distribution, Agro-Alimentaire ou téléphonie. Les entités nommées de type *+Bus* sont principalement associées à des données financières

Num. requête	Nbre total de Requêtes	Requêtes avec [DATA-TYPE]	%
Requête 1	1398	200	14,30%
Requête 2	789	181	22,68%
Requête 3	452	96	21,23%
Requête 4	263	50	19,01%
Requête 5	156	47	30,12%
Requête 6	101	25	24,75%
Requête 7	68	17	25,00%
Requête 8	44	8	18,18%
Requête 9	34	3	8,82%
Requête 10	22	5	22,72%
+ de 10 requêtes	86	19	22,09%

Tableau 6.8 – Types d'informations mentionnées dans les requêtes

comme le chiffre d'affaire, les produits associés, les nouveautés d'une marque, etc (ex : « Chiffres clés de Cetelem »).

Num. requête	Nbre total de Requêtes	Requêtes avec *+Bus*	%
Requête 1	1398	125	8,94%
Requête 2	789	74	9,27%
Requête 3	452	31	6,85%
Requête 4	263	19	7,22%
Requête 5	156	12	7,69%
Requête 6	101	5	4,95%
Requête 7	4	11	5,88%
Requête 8	44	3	6,81%
Requête 9	34	3	8,82%
Requête 10	22	3	13,63%
+ de 10 requêtes	86	8	9,30%

Tableau 6.9 – Indication de l'Entité Nommée de type *+Bus* dans les requêtes

Le tableau 6.9 montre que l'écart entre la requête 1 et les autres requêtes est beaucoup moins important que dans les autres informations relevées dans les spécificités des équations de recherche. En effet, les opérateurs booléens, la langue étrangère, les indicateurs géographiques et les types de données étaient beaucoup plus présents lors de multi-requêtages, l'entité nommée de type *+Bus* a des résultats beaucoup moins tranchés. En pourcentage, peu de requêtes contiennent autant d'entité nommée de type *+Bus* que la

première requête. Si les autres informations relevées dans les spécificités des équations de recherche semblent être des ajouts par rapport aux résultats de la première requête, il n'en est pas de même pour cet élément.

Après avoir défini les traits caractéristiques des demandes en LN ainsi que ceux de la requête, nous proposons dans la prochaine section de mettre en valeur les principales correspondances entre des différents traits.

6.3 Comparaison entre la demande en LN et la (ou les) requête(s)

Notre objectif dans cette section est de comparer les informations contenues dans les deux énoncés : à la fois de percevoir les régularités structurelles sur la façon de désigner le thème de la recherche (le référent dans la demande en LN et la requête) mais aussi de repérer éventuellement des variations syntaxiques.

Pour cela, nous avons fait une analyse en plusieurs étapes pour comparer la demande en LN et la requête :

1. Nous avons tout d'abord fait une comparaison entre la requête et le [REFERENT] de la demande en LN,

2. Puis nous avons comparé les informations contenues dans la demande en LN et celles de la (ou des) requêtes.

6.3.1 Comparaison entre le [REFERENT] de la demande en LN et la requête

La grammaire utilisée pour comparer les [REFERENTS] et les requêtes est présentée dans la sous partie 5.2.1. Elle est issue de [HUA 09] dont les travaux portent sur les comparaisons de requête dans le cadre de reformulation. Cette méthode nous permet d'appréhender cette analyse sous trois angles :

1. **une correspondance totale entre le REFERENT et la requête** *i.e.* la même forme de surface incluant des modifications mineures comme l'usage des accents ou les majuscules,

2. **une correspondance partielle entre le REFERENT et la requête** avec des variations morpho-syntaxiques comme les processus de flexion ou de dérivation, les suppressions, modifications ou ajouts de termes.

3. **une absence de correspondance** dans un premier temps au moins du point de vue surface ; les [REFERENTS] et les requêtes concernés ont alors bénéficié d'un niveau supplémentaire d'analyse à savoir celui du sens (détections des relations de synonymie, d'hypéronymie, d'hyponymie ou de métonymie à partir de thésaurus).

6.3.1.1 Correspondances totales entre le [REFERENT] de la demande en LN et la requête

Nous avons recensé 430 correspondances totales (soit 30,75 %) entre les demandes en LN et le (ou les) [REFERENT(S)] de la requête ; c'est-à-dire que le ou les terme(s) utilisé(s) étaient *stricto sensu* identiques comme nous le présentons dans les Figures 6.16 et 6.17.

Nous répertorions également dans cette section les comparaisons manifestant de modifications 'mineures' comme *(a)* la typographie (passage de minuscules aux majuscules) avec par exemple le [REFERENT] : « jardinerie » avec sa requête correspondance : « JARDINERIE », *(b)* la suppression des accents dans les requêtes avec par exemple comme [REFERENT] : « galvanisation à chaud » et comme requête : « galvanisation a chaud » ou encore *(c)* l'ajout d'un opérateur booléen (dans le sens où cet ajout ne change pas le sens de la demande) : « magasin informatique » qui devient dans la requête « magasin+informatique ». Ces modifications mineures sont au nombre de 73 (soit 5,22%). A noter que dans l'étude de [HUA 09], l'ajout ou la suppression d'accents d'une requête à une autre est comptabilisée comme une modification. Nous considérons dans notre étude que c'est une modification graphique qui peut être détectée et modifiée assez aisément puisqu'un correcteur orthographique, une correspondance avec des lexiques ou encore un calcul de similarité pourraient enrayer le problème. Nous le considérons pas comme un changement de sens mais plutôt comme un usage différencié des SRI.

Un total de 503 demandes sur 1398 (soit 35,97 %) présentent donc une correspondance totale entre le référent de la demande en LN et la requête.

6.3.1.2 Correspondances partielles entre le [REFERENT] de la demande en LN et requête

Les correspondances partielles entre le [REFERENT] et la requête sont au nombre de 446 (soit 31,90%) ; elles représentent certaines variations entre le référent de la demande en LN et la requête. Les variations résultent d'une transformation phonétique, morphologique ou syntaxique d'une suite de mots apparentés [HAB 98]. Certaines de

Figure 6.16 – Exemple 1 d'une correspondance totale entre le [REFERENT] de la demande en LN et la requête

Figure 6.17 – Exemple 2 d'une correspondance totale entre le [REFERENT] de la demande en LN et la requête

ces variations constituent des *variantes, i.e* des équivalents effectifs de l'expression en cause (*infractus myocardique* pour *infractus du myocarde* par exemple). Nous présentons dans cette sous-section les résultats de nos analyses sur ces variations.

[A] Variations morpho-syntaxiques entre le [REFERENT] de la demande en LN et la requête : nous avons relevé ici certaines transformations flexionnelles et dérivationnelles (voir à ce sujet la sous partie 3.3.2.1 à la page 43). Nous avons pour cela utilisé

Figure 6.18 – Exemple d'une correspondance totale entre le bloc [PRECISIONS] de la demande en LN et la requête

l'analyseur syntaxique *Xelda*.

- **les processus de flexion** *i.e.* procédé de modification de la forme de référence d'un mot nommée *forme canonique* ou *lemme*, *i.e* la forme d'un mot sans ses marques dites de *flexions*. Il s'agit des flexions de genre, de nombre pour les noms, les adjectifs et les pronoms et des flexions de personne, de nombre, de temps, de mode et de voix pour les verbes. Nous étudions ici principalement les processus de flexion pour les noms et plus précisément au travers des marques de nombre (singulier => pluriel ou pluriel => singulier) car ce sont les plus nombreux dans notre corpus : 83 demandes (soit 5,93%) en LN présentent une variation de nombre. Dans la demande en LN « Je désire trouver une étude de marché sur l'automobile », le nom au singulier du [REFERENT] « automobile » subit une transformation du nombre au moment de la requête : « automobiles ». Les autres processus de flexion, notamment celui au travers des marques de genre sont inexistants dans notre corpus et ne sont donc pas répertoriées.

- **les processus de dérivation** *i.e.* procédé de formation de mots nouveaux par modification (addition, suppression ou remplacement) d'un morphème (suffixe) par rapport à une base (radical) : 48 demandes (soit 3,43%) en LN présentent un processus de dérivation . On peut avoir des modification d'un nom, d'un verbe, ou encore d'un adjectif. Dans la demande en LN « je recherche les dernières études

dans le domaine bancaire » : « bancaire » se transforme en « banque » dans la requête.

[B] Modifications de la langue : certains besoins informationnels sont formulés en français dans les demandes en LN et en anglais dans les requêtes. Cela concerne 16 requêtes (soir 1,14 %) des requêtes 1. Exemple : « maquillage » dans le [REFERENT] et « make-up » dans la requête. Le [REFERENT] ne présente pas de termes en anglais.

[C] Modification(s) d'un ou plusieurs termes du [REFERENT] de la demande LN : cette transformation concerne un grand nombre de requêtes puisque 299 (soit 21,38%) d'entre-elles conservent des éléments du [REFERENT] de la demande en LN mais avec des transformations. Nous avons répertorié trois types de transformations : *(C.1)* les ajouts de termes (concerne 55 demandes), *(C.2)* les glissements de termes (concerne 43 demandes) et *(C.3)* les suppressions de termes (concernent 201 demandes).

[C.1] Premier type de transformations entre le [REFERENT] et la requête : les ajouts de termes : 55 requêtes (soit 3,93%) conservent des éléments du [REFERENT] mais se voient ajouter des termes. Ces ajouts sont majoritairement des NOUN (ex : [REFERENT] : « bijouterie » => requête : « horlogerie bijouterie » ou encore [REFERENT] : « piscines » => requête : « protection piscines ») et des ADJ ([REFERENT] « nettoyage » => requête : « nettoyage industriel » ou encore [REFERENT] : « produits naturels » => requête : « produits pharmaceutiques naturels »). Le détail des ajouts de termes dans les requêtes sont décrits dans le Tableau 6.10.

Type d'ajouts	Nombre
Ajout d'un ou plusieurs NOUN(s)	32
Ajout d'un ou plusieurs ADJ(s)	10
Ajout d'un NOUN [DE] NOUN	4
Ajout d'un +WH (Interrogatif)	3
Ajout d'une PREP (pour former un NOUN [DE] NOUN	3
Ajout d'une PROP	3

Tableau 6.10 – Analyse du premier type de transformation : les ajouts de termes

[C.2] Deuxième type de transformations entre le REFERENT et la requête : les glissements de termes Nous avons relevé deux types de glissements :

1. **Glissements de termes dans une structure grammaticale équivalente** : ce glissement de termes n'altère pas la catégorie grammaticale entre le [REFERENT] et la requête. Il concerne 29 requêtes et s'applique aux structures NOUN [DE] NOUN (pour 23 d'entre elles) (ex : [REFERENT] : « affinage d'aluminium » => requête : « recyclage d'aluminium ») et les structures NOUN ADJ (exemple : [REFERENT] « veille technologique » => requête : « veille stratégique ») pour les 6 autres requêtes.

2. **Glissements de termes dans une structure grammaticale différente** : on conserve le même nombre de termes entre le [REFERENT] et la requête mais la catégorie grammaticale d'un ou plusieurs termes de l'expression change (exemple : [REFERENT] : « compléments alimentaires minceur » => requête : « compléments alimentaires Brésil », le nom « minceur » du [REFERENT] est devenu le nom propre « Brésil » de la requête) .

Ces glissements sont au nombre de 12 (soit 0,85%) et sont détaillés dans le Tableau 6.11.

Type de glissements	Nombre
Glissement d'un NOUN en un ADJ	6
Glissement d'un VB en un NOUN	4
Glissement d'un ADJ en un NOUN	2
Glissement d'un NOUN en une PROP	2

Tableau 6.11 – Analyse du deuxième type de transformations : les glissements de termes avec modification de la structure grammaticale

Le nombre total des glissements de termes est donc de 43 soit 3,07%.

[C.3] Troisième type de transformation entre le REFERENT et la requête : les suppressions de termes Enfin, un troisième type de transformations entre le [REFERENT] et la requête est la suppression de termes. Cette transformation est la plus importante puisqu'elle concerne 201 requêtes (soit 14,37%) (exemples : [REFERENT] : « boulangerie traditionnelle » => requête : « boulangerie » ou encore [REFERENT] : « organisateurs de salons » => requête : « salons ») . Les principaux types de suppressions sont présentés dans le Tableau 6.12 avec leurs occurrences en quatre partie : *(a)* suppression d'un terme comme un NOUN, un ADJ, une PROP ou une COORD, suppression de plusieurs termes

comme un NOUN ADJ ou un NOUN [DE] NOUN. *(b)* suppression d'une partie des termes dans les expressions NOUN [DE] NOUN *(c)* suppression d'un des terme dans l'expression NOUN ADJ *(d)* suppression d'un des termes dans les deux tri-grammes NOUN ADJ ADJ et NOUN ADJ NOUN.

Type de suppressions	Nombre
Suppression d'un ou plusieurs NOUN => Ø	53
Suppression d'un ou plusieurs ADJ => Ø	41
Suppression d'une PROP => Ø	7
Suppression d'un NOUN [DE] NOUN => Ø	5
Suppression d'un NOUN ADJ => Ø	2
Suppression d'une COORD => Ø	1
Passage d'un NOUN [DE] NOUN => NOUN	55
Passage d'un NOUN [DE] NOUN => NOUN ADJ	2
Passage d'un NOUN [DE] NOUN => NOUN NOUN	2
Passage d'un NOUN [DE] NOUN => ADJ	1
Passage d'un NOUN ADJ => NOUN	18
Passage d'un NOUN ADJ => ADJ	3
Passage d'un NOUN ADJ ADJ => NOUN ADJ	8
Passage d'un NOUN ADJ NOUN => NOUN ADJ	3

Tableau 6.12 – Analyse des principaux schémas de suppressions de termes

D'après le tableau 6.12, nous pouvons constater que les modifications les plus importantes entre les formulations des besoins en LN *via* le REFERENT ou *via* la requête portent sur la suppression d'un ou plusieurs NOUN et dans une moindre mesure sur la suppression d'un ou plusieurs ADJ.

6.3.1.3 Absence de correspondance entre le [REFERENT] de la demande en LN et la requête

Dans cette sous partie, nous présentons les demandes en LN dont les [REFERENTS] et les requêtes n'ont pas eu de correspondance entre eux. Ils sont au nombre de 254 (soit 18,16%).

Nous avons eu besoin de recourir à un réseau sémantique pour l'exploitation des traits basés sur les liens sémantiques. Pour cela, nous avons exploité le thésaurus interne utilisé par le moteur de recherche `Plusdetudes.com` composé de 180 000 termes pour la catégorisation automatique des documents.

Nous avons pu ainsi définir plusieurs types de relations sémantiques entre les termes des [REFERENTS] des demandes en LN et les termes des requêtes :

1. **environnement sémantique proche (synonymie)** : les termes utilisés entre la demande en LN et la requête sont des synonymes (« habit » => « vêtement ») ou des termes issus de la même branche du thésaurus (« sécurité » => « incendie »). Cela concerne 119 demandes (soit 8,51%). Ce chiffre assez important est prometteur ; il indique qu'en se basant sur les demandes des utilisateurs en LN pour formuler les besoins informationnels, les SRI pourraient s'appuyer sur davantage de termes (*i.e.* issus de la demande en LN) que ceux initialement utilisés dans les barres de recherche par les utilisateurs pour formulés leurs recherches. L'apport serait alors d'ajouter des termes sémantiquement proches de la requête initiale pour la désambiguïser. C'est d'autant plus vrai pour les requêtes uni-grammes qui souffrent de manque contextualisation et qui concernent 58 de ces 119 demandes.

2. **relation d'hyperonymie** : relation sémantique hiérarchique d'un terme à un autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus spécifique. Cela concerne 41 demandes (soit 2,93%). Exemple : « peinture » => « art »

3. **relation d'hyponymie** : relation sémantique d'un terme à un autre selon laquelle l'extension du premier est incluse dans l'extension du second. Cela concerne 13 demandes (soit 0,92%). Ex : « légumes » => « oignons »

4. **relation de métonymie** : relation sémantique qui consiste à remplacer le terme propre par un autre qui lui est proche ou qui en représente une qualité (ici,toutes les entités nommées qui étaient utilisées pour mentionner un objet) et qui a avec lui une relation logique. Ex : « automobile » => « Mercedes ». Cela concerne 40 demandes (soit 2,86%).

5. **environnement sémantique inexistant** : malgré le recours au thésaurus, il n'y a pas eu de relation établie entre les termes. Cela concerne 41 des demandes (soit 2,93%). Ex : « norme DVB/RCS » => « satellite ». Cela n'exclue pas le fait qu'ils aient une relation mais simplement que cette relation n'a pas été identifiée à l'aide des ressources lexicales. Ceci peut s'expliquer notamment par l'absence des termes comme « norme DVB/RCS » dans les thésaurus.

En ce qui concerne l'absence de correspondance, nous pensons qu'il s'agit d'une faiblesse du thésaurus puisque l'analyse des autres blocs d'informations ne permet pas d'améliorer ce résultat. Après ce recoupement avec une ressource externe et cette analyse, nous pourrions donc comptabiliser les résultats des [REFERENTS] et des requêtes issus des relations de synonymie, d'hyperonymie, d'hyponymie et de métonymie soit 213 demandes (15,24%) comme ayant une relation sémantique. Ces résultats pourraient ainsi être comptabilisés dans la partie 'correspondances partielles' et non dans la partie 'Absence de correspondance'. Il reste à ce stade seulement 41 demandes pour lesquelles nous n'avons pas pu corroborer de correspondances entre la demande en LN et la requête.

6.3.2 Comparaison entre les autres blocs d'informations de la demande en LN et la requête

Nous nous sommes basés principalement sur le bloc [REFERENT] pour établir les comparaisons avec la requête. Mais l'avantage de la demande en LN est de pouvoir recueillir les informations sous différentes formes et à différents « emplacements » de la demande. Cette richesse fait aussi la complexité de l'analyse. Nous recueillons ici les analyses des blocs [TYPE-DONNÉES], [PRÉCISIONS] et [CONTEXT] qui peuvent s'avérer également riches car porteurs d'informations.

Correspondances totales entre les autres blocs de la demande en LN et la requête Nous avons également comptabilisé les autres correspondances situées non plus dans le [REFERENT] mais dans les autres blocs d'informations de la demande en LN à savoir [TYPE-DONNÉES], [PRÉCISIONS] et [CONTEXTE]. Ainsi, 71 demandes sur les 1398 (soit 5,07%) présentent une correspondance totale avec les blocs [TYPE-DONNÉES], [PRÉCISIONS] ou [CONTEXTE]. Nous donnons un exemple dans la Figure 6.18.

Correspondances partielles entre les autres blocs de la demande en LN et la requête :

Il s'agit des correspondances qui présentent certaines variances (présentées dans 6.3.1.2) entre les blocs [TYPE-DONNÉES], [PRÉCISIONS] et [CONTEXTE] de la demande en LN et la requête. Elles sont au nombre de 65 soit 4,64 %.

Enfin, si nous comparons les autres éléments compris dans la demande en LN autres que le [REFERENT], nous recoupons 136 demandes. On peut ainsi conclure que *(a)* la requête contient des informations qui peuvent être plus éparses quand il s'agit de le

formuler en LN, *(b)* notre segmentation en blocs d'information est trop stricte et demanderait à être étendue aux blocs [PRÉCISIONS] et [CONTEXTE] notamment afin d'obtenir un maximum d'informations (ou tout simplement envisager de les prendre en compte de façon plus adéquate pour une analyse plus performance des demandes en LN).

Conclusion : Comparaison des informations contenues dans la demande en LN et la requête

Dans cette dernière section, nous proposons une synthèse des principales caractéristiques entre les demandes en LN et les requêtes. Dans le Tableau 6.13, nous représentons en pourcentages les principales informations contenues dans les deux énoncés.

Types d'informations	LN	Requêtes
Géographie	34,19%	8,51%
Date	3,43%	0,14%
Prix	2,57%	Ø
Types de données	88,84%	19,02%
Délai	1,07%	Ø
Marque / nom de société (+Bus)	10,87%	8,38%

Tableau 6.13 – Synthèse des informations contenues dans la demande en LN et dans la requête

On peut remarquer que les deux types d'informations qui sont présentes dans les demandes en LN et assez peu dans les requêtes sont la géographie et le types de données. Ces informations sont pourtant importantes pour sélectionner une étude de marché.
D'autres informations comme la date, le prix et le délai souffrent également de lacunes lors du passage de la demande en LN à la requête mais de façon moins prononcée. Les noms de marque / noms de société *+Bus* sont assez bien représentées dans les requêtes ; c'est le type d'informations qui semble le moins souffrir du passage de la demande en LN à la requête.

Nous remarquons à partir du Tableau 6.14 que les proportions de NOUN, PROP et ADJ dans les uni-grammes sont en moyenne équivalentes entre le [REFERENT] de la demande en LN et la requête. Un écart reste important entre les abréviations, acronymes et sigles : ils sont en effet plus importants dans les requêtes ce qui peut créer des problèmes d'ambiguïté (car non contextualisés). Les NOUN (aussi bien dans les bi-grammes que dans les tri-grammes) sont plus importants dans les requêtes, au détriment d'autres

N-grammes	Catégories morpho-syntaxique	[REFERENT]	Requêtes
Uni-grammes	NOUN	77,48%	76,08%
Uni-grammes	PROP	11,44%	13,93%
Uni-grammes	ADJ	5,40%	6,84%
Uni-grammes	ABBR|ACCR|SIGLES	0,02%	5,36%
Bi-grammes	NOUN ADJ|ADJ NOUN	70,02%	41,87%
Bi-grammes	NOUN NOUN	16,48%	33,21 %
Bi-grammes	NOUN PROP|PROP NOUN	3,20%	9,38%
Tri-grammes	NOUN PREP NOUN	77,63%	48,34%
Tri-grammes	NOUN ADJ NOUN	8,07%	2,80%
Tri-grammes	NOUN ADJ ADJ	8,07%	2,81%
Tri-grammes	NOUN NOUN NOUN	0,57%	7,92%
Quadri-grammes	NOUN PREP NOUN + 1	68,57%	55,74 %
Quadri-grammes	NOUN ADJ NOUN + 1	19,04%	7,09%
Quadri-grammes	PROP PROP PROP +1	0,95%	Ø
Quadri-grammes	NOUN NOUN NOUN + 1	8,57%	0,14%

Tableau 6.14 – Catégories morpho-syntaxiques principalement utilisées dans les demandes en LN et dans les requêtes

formes comme NOUN ADJ dans les bi-grammes ou NOUN [DE] NOUN dans les tri- et quardi- grammes des [REFERENTS] des demandes en LN. Des informations qualificatives intéressantes lors des requêtes comme des adjectifs affinent la recherche peuvent alors être perdues.

D'après nos analyses, les expressions NOUN ADJ et NOUN [DE] NOUN sont plus précises ; elles correspondent aux expressions communément utilisées dans les rapports économiques pour décrire un secteur d'activité (ce sont principalement ces formes qui sont développées dans les thésaurus internes de la société). Sur une recherche *full-text*, ces expressions rappellent davantage de documents que les expressions NOUN NOUN et NOUN NOUN NOUN qui sont moins présentes dans les documents de type étude de marché. Une constatation-clef à ce stade est que les expressions utilisées dans les requêtes sont potentiellement des formes qui donneront moins de résultats que les expressions utilisées dans les demandes en LN.

La longueur moyenne des requêtes est de 2,40 termes alors qu'elle est un tout petit peu plus longue dans les [REFERENTS] : 2,52. Les termes utilisés à la fois dans les [REFERENTS] et les requêtes sont soit parfaitement équivalents dans 41,1% des cas soit ayant subi une modification mineure (processus de flexion ou de dérivation sur l'un

des termes, ajout ou suppression d'un ou plusieurs termes dans une expression) dans 40,7% des cas. Dans 10,8% des cas, on peut faire un rapprochement sémantique entre le [REFERENT] et la requête (relations de synonymie, d'hyperonymie, d'hyponymie ou de métonymie). Enfin, dans 2,00% des cas, nos analyses n'ont pas permis de trouver de relation sémantique entre le [REFERENT] et la requête. Une hypothèse serait que les utilisateurs pensent que les formulaires en SAV dans lesquels ils saisissent leurs demandes en LN et les formulaires d'interrogations dans lesquelles ils entrent leurs requêtes sont des champs qui se complètent et donc n'ont pas besoin de répéter le thème initial.

Très peu d'utilisateurs modifient l'ordre d'apparition des termes ou des expressions entre le [REFERENT] et la requête : seulement 24 requêtes (soit 1,71%) indiquent un inversement de l'ordre entre les mots de la demande en LN et les termes de la requête ; 8 sont dus à un passage d'une forme adjectivale au nom propre du pays (ex : « café en France » => « marché français du café » ou encore « poisson surgelé en Espagne » => « marché espagnol du poisson surgelé ». Nous observons qu'un ajout de type de données « marché » peut avoir lieu. Nos résultats sont en accord avec [BRU 97]. Leur étude portait sur des reformulations effectuées par les utilisateurs eux-mêmes : ils ont montré que la reformulation consiste souvent en *(i)* la répétition de la requête initiale, *(ii)* l'augmentation ou la diminution de quelques mots, *(iii)* le changement de l'orthographe de la requête, *(iv)* l'utilisation de formes dérivées ou d'abréviations. Par ailleurs, les utilisateurs associent parfois les mots synonymes ou des concepts voisins (par rapport au domaine visé par la requête initiale). Également, dans une étude sur `Dogpile.com`, [JAN 07] reporte que 37 % de leurs requêtes sont des reformulations. C'est également le cas dans 27 % de l'étude de [PAS 06] ou dans 28 % des *logs* d'AOL de l'étude de [HUA 09].

Pour conclure, nos résultats indiquent peu de différences sur le vocabulaire utilisé entre les deux types d'énoncés ; seuls quelques ajouts et suppressions de termes ont été modifiés . Les expressions dans les [REFERENTS] peuvent éventuellement être considérées comme plus adéquates pour la recherche car plus précises. Il n'existe pas d'apports majeurs à utiliser les demandes en LN de ce coté là. En revanche, nous observons un apport indéniable dans la demande en LN dans les blocs [CONTEXT] et [PRECISIONS] qui permettent d'obtenir deux informations complémentaires très intéressantes : la zone géographique et le type de données économiques désirés. Ces aspects présentent le plus de différence entre les demandes en LN et les requêtes. Bien que connexes, ces informations sont riches principalement pour contextualiser le besoin informationnel et plus spécifiquement pour déterminer les critères de sélection d'études de marché.

Dans le chapitre suivant, nous évaluerons si des caractéristiques entre la demande en LN et la requête se distinguent en fonction des buts des tâches de RI. Nous proposerons alors un premier ensemble de résultats qui pourront améliorer notre compréhension du comportement et des besoins des utilisateurs. Ces observations nous amèneront enfin à proposer un ensemble de recommandations en conclusion pour améliorer le système de recherche d'informations.

CHAPITRE 7

QUELLE EXPRESSION DES BESOINS INFORMATIONNELS SELON LA TÂCHE DE RI ?

Nous souhaitons dans ce chapitre comparer les différents résultats présentés dans le chapitre précédent en opérant plus spécifiquement une distinction entre les différents groupes d'utilisateurs identifiés dans la section 5.1.3 en fonction des objectifs de RI.

Pour rappel, la typologie des types de tâches de RI est la suivante :

- [TACHE-CREA] : **le but de la tâche de RI est la création d'entreprise ou le lancement d'un nouveau produit ou encore d'une nouvelle marque** ; les objectifs opérationnels sont de se procurer une étude de faisabilité, d'identifier la concurrence éventuelle,
- [TACHE-SCO] : **le but de la tâche de RI est la réalisation d'une tâche scolaire** ; les objectifs opérationnels sont de préparer un examen, d'écrire un mémoire, de travailler sur une étude de cas…
- [TACHE-PRO] : **le but de la tâche de RI est l'obtention d'informations dans un cadre professionnel** ; les objectifs opérationnels sont de mieux connaitre le marché, ses éléments chiffrés, d'identifier les tendances d'un marché ou d'un produit, de faire de la veille stratégique…

Ce chapitre permettra de mettre en évidence les écarts de comportement le cas échéant entre ces trois groupes d'utilisateurs. Il reprend en partie les résultats présentés dans [LAT 13a].

7.1 Présentation générale des demandes en LN différenciées selon les types d'utilisateurs

Les résultats seront présentés selon trois axes :

1. les traits morphologiques de la demande : principales caractéristiques des demandes en LN afin de dégager d'éventuelles régularités comme la longueur, l'usage des noms de marque, des noms de pays, des dates, etc.

2. les traits syntaxiques de la demande : la structure syntaxique des demandes, les pronoms personnels, les types de phrases employées, etc.

3. les traits sémantiques de la demande : le nombre d'éléments polysémiques dans la demande, le complexité linguistique (profondeur des nœuds de le thésaurus), etc.

7.1.1 Traits morphologiques de la demande en LN différenciés par type de tâche en RI

En ce qui concerne les traits morphologiques, nous avons pu relever certaines différences entre la formulation des demandes en LN et la tâche de RI. La [TACHE-PRO] formule avec moins de termes ses besoins informationnels : en moyenne 17,66 termes contre 21,51 pour la [TACHE-CREA] et 22,27 pour la [TACHE-SCO]. Cette distribution est représentée dans les Figures 7.1 et 7.2.

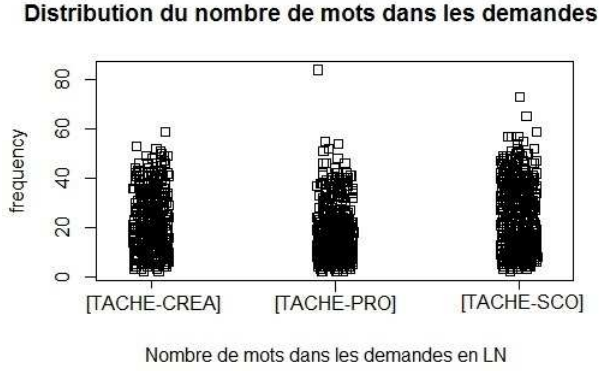

Figure 7.1 – Distribution du nombre de mots dans la demande en LN par type de tâche de RI

Les histogrammes de la figure 7.2 différencient la distribution de la [TACHE-PRO] (histogramme de gauche) ; les deux premières catégories représentent plus de 150 demandes en LN qui contiennent moins de 20 termes. Peu de demandes sont longues dans ce groupe de personnes.

En ce qui concerne l'usage et la répartition des concepts à l'intérieur de la demande en LN, plusieurs caractéristiques se dégagent notamment à partir de la figure 7.3. Les deux concepts qui subissent le plus de variations entre les différents groupes d'utilisateurs sont la [FONCTION-CLIENT] et le [CONTEXTE]. En effet, le groupe [TACHE-SCO]

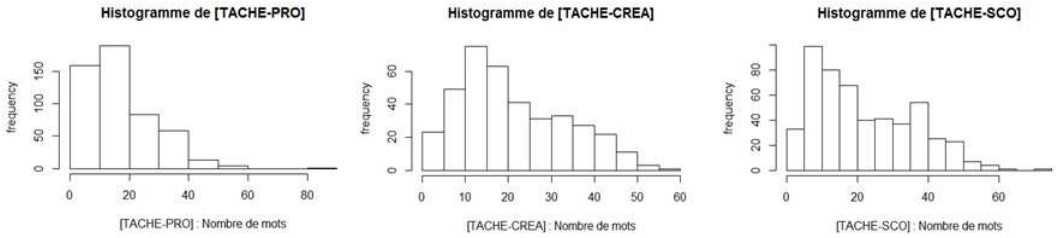

Figure 7.2 – Distribution du nombre de mots dans la demande en LN par type de tâche de RI, représentation en histogrammes

utilise d'avantage le concept [FONCTION-CLIENT] pour indiquer qu'ils sont étudiants et éventuellement précise le degré d'étude (master, licence) ou encore le nom de leur université. Le [CONTEXTE] est très peu utilisé par le groupe [TACHE-PRO] par rapport aux deux autres groupes d'utilisateurs : très peu d'informations supplémentaires sont données en plus du secteur recherché qui permettraient de mieux contextualiser leurs demandes avec des exemples ou en définissant plus précisément le cadre de leur recherche. Cette conclusion corrobore notamment le fait que le groupe tend à utiliser moins de termes pour formuler ses demandes en LN.

De façon moins prononcée, le concept [SALUTATION] est moins représenté par la [TACHE-PRO]. Cela suggère une démarche plus rapide et plus directe du formulaire SAV. Le groupe [TACHE-CREA] mentionne moins souvent que les deux autres groupes d'utilisateurs le type d'informations souhaité [TYPE-DONNÉES].

Nous relevons également d'autres écarts de comportements en fonction des types de tâche de RI sur les concepts de prix, de dates, de marques ou de géographies.

Ces concepts sont repris dans la Figure 7.4 à partir de laquelle nous observons que la géographie a une place plus importante pour le groupe d'utilisateurs ayant une [TACHE-CREA] ; ceci s'explique principalement par le fait que le développement d'une activité professionnelle est très liée à son emplacement géographique (*e.g.* « ouvrir un restaurant à Bordeaux »). C'est le concept principalement développé dans les demandes en LN de ce groupe d'utilisateur : les concepts de dates, marques et prix apparaissent alors secondaires dans la façon de présenter leurs besoins. Le concept de prix est prédominant pour la [TACHE-SCO] qui demande une réduction voire la gratuité des études. Rappelons que la [TACHE-SCO] regroupe les utilisateurs qui ont une tâche scolaire à effectuer. Ce concept est moins présent pour la [TACHE-CREA] et inexistant pour la [TACHE-PRO]. Le concept de marque est également important pour la [TACHE-SCO] : les demandes com-

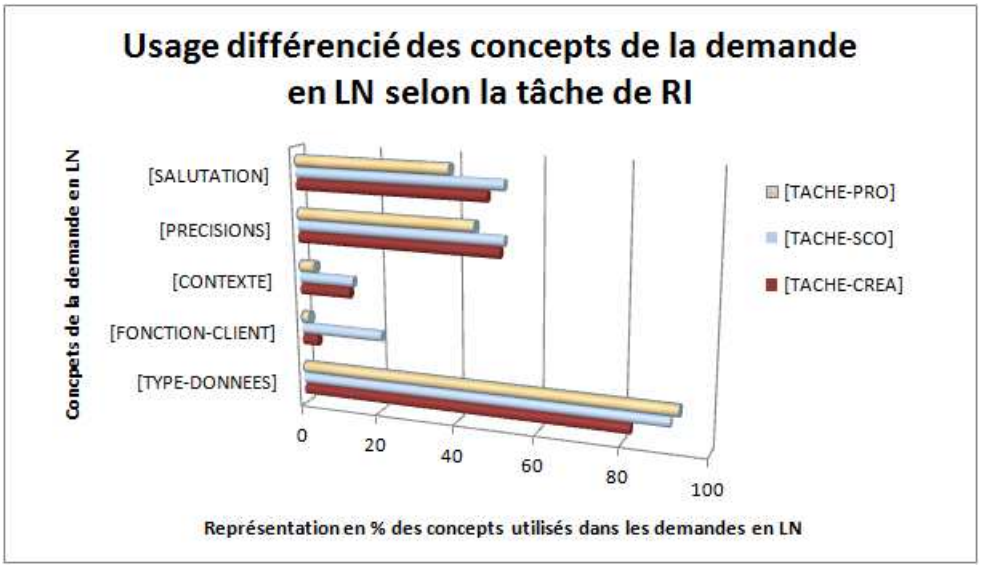

Figure 7.3 – Usage différencié des concepts de la demande en LN selon la tâche de RI

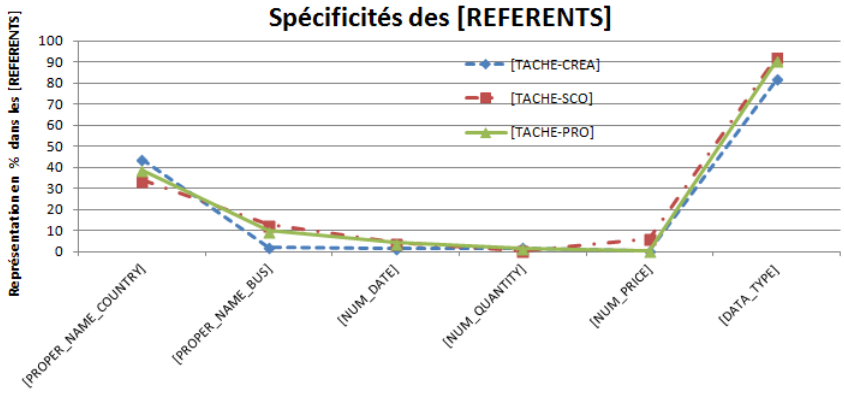

Figure 7.4 – Spécificités des [REFERENTS] dans les demandes en LN par type de tâche de RI

portent alors de nombreuses mentions de noms de marque ou de sociétés qui peuvent être l'objet même de leur travail (« réaliser une étude sur Coca-Cola ») ou sur un marché mais avec une demande particulière sur les principaux acteurs du secteur (*i.e* « les principaux leaders du marché des boissons énergétiques (CA de Isostar) »). Le concept de dates est utilisé de façon similaire par les groupes [TACHE-SCO] et [TACHE-PRO] ; ces utilisateurs ayant des contraintes de temps (délai) ou des demandes plus précises sur le scope temporel que doivent couvrir les études de marché. Les [TYPES-DONNEES] (non

représentées dans le schéma) sont de 81,79% [TACHE-CREA], 92,39% pour la [TACHE-SCO] et 90,51% [TACHE-PRO].

7.1.2 Traits syntaxiques de la demande en LN différenciés par type de tâche en RI

Les principaux traits syntaxiques étudiés sont présentés dans la Figure 7.5. Notons tout d'abord que le groupe d'utilisateurs [TACHE-PRO] utilise davantage les tournures impersonnelles et moins de pronoms personnels (PP) que les deux autres groupes. Le groupe [TACHE-SCO] emploie plus souvent la première personne du pluriel, se situant dans une démarche de groupe (« nous effectuons un dossier sur »). Pour sa part, le groupe sur la [TACHE-CREA] utilise davantage la première personne du singulier, se situant dans une démarche plus personnelle (« je vais créer une entreprise »). Nous notons également que les utilisateurs ayant une [TACHE-SCO] s'expriment plus avec des phrases syntaxiquement correctes, notamment par rapport aux utilisateurs ayant une [TACHE-PRO] qui eux utilisent souvent une syntaxe incorrecte avec des phrases partielles et/ou incomplètes (exemple : « étudier la croissance de Novartis »). Certaines de leurs demandes en LN peuvent s'apparenter d'ailleurs à des requêtes (exemple : « marché du parquet »).

Il semblerait que le groupe d'utilisateurs [TACHE-PRO] formule ses besoins en LN dans une structure moins formelle que les deux autres groupes ; leurs demandes en LN se rapprochent d'avantage à une formulation hybride entre une expression en LN et une requête.

7.1.3 Traits morpho-syntaxiques du [REFERENT] de la demande en LN selon les types d'utilisateurs

A partir du découpage en blocs d'informations, nous avons effectué un travail plus spécifique sur les référents. En effet, comme nous l'avons vu à la sous-section 5.1.2.2 (page 80), c'est dans ce bloc d'informations (qui peut contenir de 1 à 7 référents) que se retrouvent la plupart des éléments également formulés dans la requête du moteur de recherche. Le [REFERENT] est le concept porteur de l'information principale de la demande en LN ; il contient le thème de la recherche (secteur d'activité recherché). Afin de pouvoir être comparé plus finement avec la requête, ce bloc a fait l'objet d'une analyse qualitative avec une étude morpho-syntaxique de tous ses termes constitutifs.

Le nombre de [REFERENTS] dans les demandes en LN par type de tâche de RI : ce nombre est présenté dans la Figure 7.6. Cette figure représente la distribution du

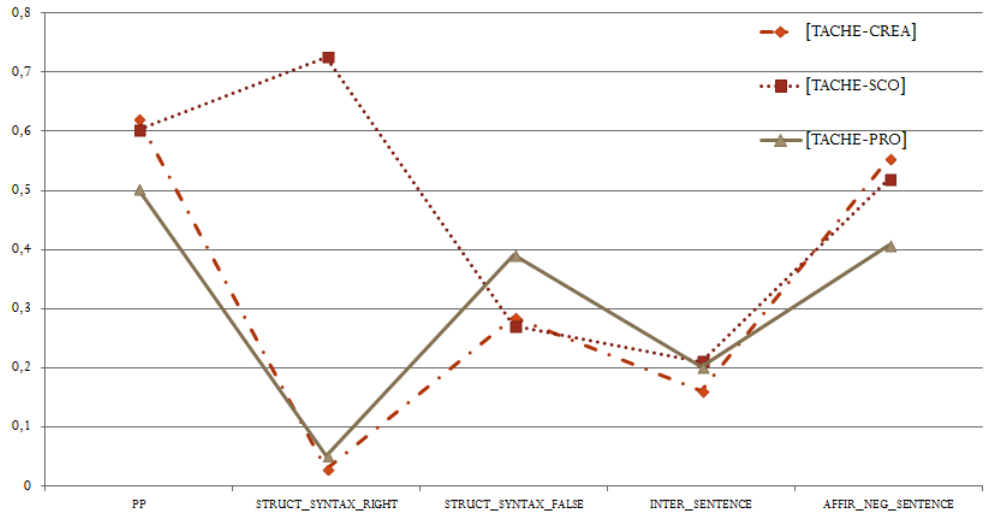

Figure 7.5 – Traits syntaxiques des demandes en LN différenciés par type de tâche de RI

nombre de [REFERENTS] ; allant de $R = 0$ référent à $R = 7$ référents dans la demande en LN). Nous déduisons de cette figure que le groupe [TACHE-SCO], représenté en pointillé dans le schéma, formule davantage ses besoins informationnels avec un seul référent : 84,80% de leur demandes se verbalisent dans le [REFERENT-1] contre 67,28% pour le groupe [TACHE-CREA] et 65,61% pour le groupe [TACHE-PRO]. Les groupes d'utilisateurs [TACHE-CREA] et [TACHE-PRO] semblent avoir le même comportement par rapport au nombre de [REFERENTS] dans les demandes en LN.

La longueur des n-grammes des [REFERENTS] : cette longueur est présentée dans la Figure 7.7. La longueur des n-grammes dans les [REFERENTS] est de 1 terme ($n = 1$) dans 49,32% pour l'ensemble des [REFERENTS] du groupe [TACHE-SCO], 47,42% pour le groupe [TACHE-PRO] et de 38,92% pour la [TACHE-CREA]. Ce dernier groupe apparait en pointillé sur le graphique ; la distribution de la longueur des n-grammes est différente en donnant plus de poids aux formulations longues. Le $n = 0$ sur la Figure représente les 8 demandes en LN ne contenant pas de [REFERENT].

L'analyse morpho-syntaxique des [REFERENTS] : l'analyse morpho-syntaxique du bloc [REFERENT] de la demande en LN (voir aussi la partie 5.1.2.2, page 80) a été différenciée selon la tâche de RI des utilisateurs. Elle a relevé un usage différencié des catégories morpho-syntaxiques que nous présentons dans les Figures 7.8 et 7.9.

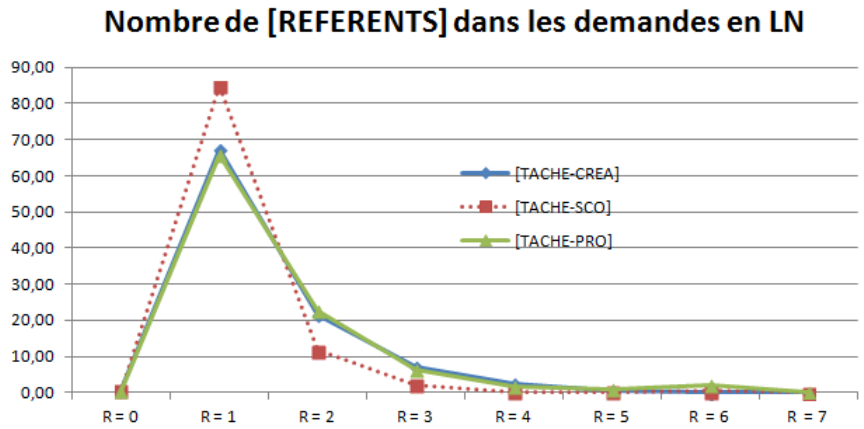

Figure 7.6 – Distribution du nombre de [REFERENTS] dans les demandes en LN selon le type de tâche de RI

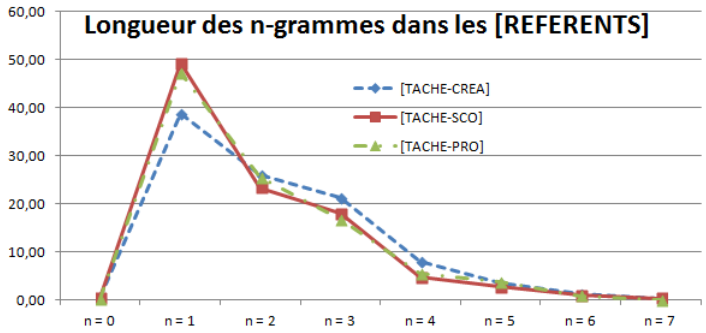

Figure 7.7 – Longueur des n-grammes dans les [REFERENTS] différenciée par type de tâche de RI

Il se dégage de ces graphiques certains traits caractéristiques récurrents par typologie d'utilisateurs :

- la [TACHE-CREA] utilise des noms et principalement des noms au singulier (NOUN-SG) dans le cas d'uni-grammes ainsi que des noms propres (PROP) et très peu d'autres formes, des NOUN NOUN dans le cas de bi-grammes ainsi que des NOUN ADJ, des NOUN PREP NOUN pour les tri- et quadri-grammes ainsi que des expressions composées exclusivement de NOUN,
- la [TACHE-SCO] : utilise davantage les noms propres (PROP) à la fois dans les uni- et bi-grammes. Ces utilisateurs ont ensuite tendance dans les syntagmes plus longs

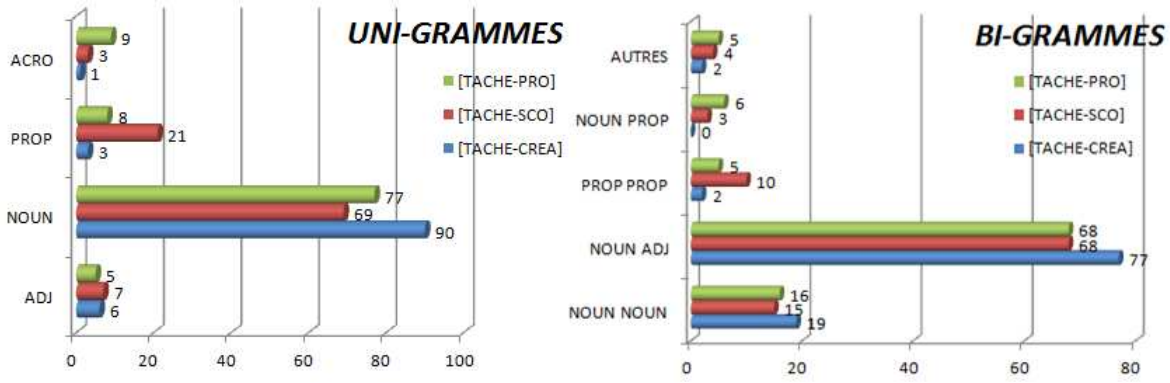

Figure 7.8 – Catégories morpho-syntaxiques pour les uni- et bi-grammes des [REFERENTS] issus des demandes en LN et différenciés par type de tâche de RI (en pourcentages)

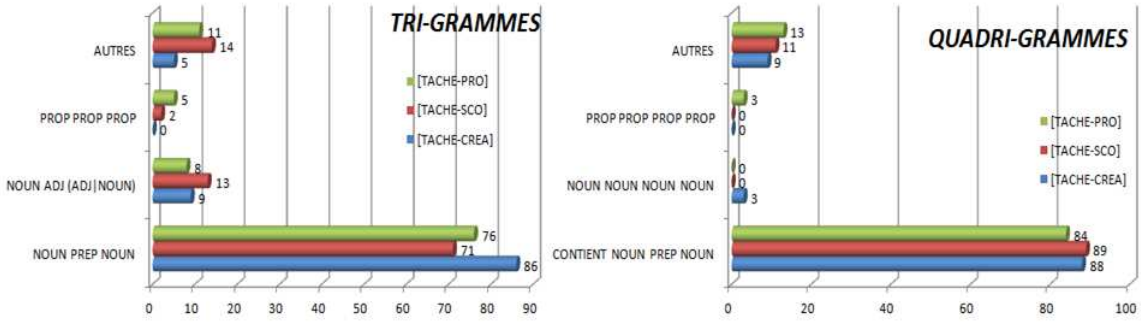

Figure 7.9 – Catégories morpho-syntaxiques pour les tri- et quadri-grammes des [REFERENTS] issus des demandes en LN et différenciés par type de tâche de RI

à introduire une forme adjectivale pour formuler les référents,

- la [TACHE-PRO] a davantage recours à des acronymes, ainsi qu'à des noms pour les uni-grammes, des NOUN PROP pour les bi-grammes, des PROP PROP PROP pour les tri-grammes ou encore PROP PROP PROP PROP pour les quadri-grammes.

Nous résumons ces caractéristiques dans le Tableau 7.1.

7.1.4 Traits sémantiques de la demande en LN différenciés par type de tâche en RI

Nous avons mesuré plusieurs aspects dans les traits sémantiques de la demande en LN : *(a)* la valeur polysémique [POLYSEMY-VALUE] : nombre de fois que le terme ou

	Uni-grammes	Bi-grammes	Tri-grammes	Quadri-grammes
[TACHE-CREA]	NOUN PROP	NOUN NOUN NOUN ADJ	NOUN PREP NOUN NOUN NOUN NOUN	NOUN PREP NOUN +1 NOUN NOUN NOUN NOUN
[TACHE-SCO]	PROP	PROP PROP	NOUN ADJ ADJ NOUN ADJ NOUN	NOUN PREP NOUN +1
[TACHE-PRO]	ACCRONYM NOUN	NOUN PROP	PROP PROP PROP	PROP PROP PROP PROP

Tableau 7.1 – Catégories morpho-syntaxiques des [REFERENTS] issus des demandes en LN et différenciés par type de tâche de RI

les termes du [REFERENT] apparaissant dans les thésaurus, *(b)* la complexité linguistique [LINGUISTIC-COMPLEXITY] : profondeur des nœuds dans le thésaurus du [REFERENT], *(c)* l'évaluation de l'ambigüité de la tâche : calcul en fonction des traits sémantiques évaluées comme favorisant l'ambiguïté de la tâche, *(d)* les secteurs d'activité mentionnés dans les [REFERENTS] : nombre de correspondances avec les thésaurus internes de l'entreprise.

(a) La valeur polysémique [POLYSEMY-VALUE] : cette valeur est calculée en fonction du nombre de fois que les termes apparaissent sous leurs formes lemmatisées dans les thésaurus internes de la société. Plus un terme apparait dans les thésaurus, plus sa valeur polysémique est élevée car non spécifique à un secteur donné particulier. Ainsi, la *polysemy value* à 0 indique que le terme n'est mentionné qu'une seule fois dans les différents thésaurus ; il est donc peu polysémique. Les *polysemy values* 1 à 3 sont présentées dans la Figure 6.2. Les termes dont la *polysemy value* vaut $k (k \geq 0)$ apparaissent $k+1$ fois sauf si $k = 3$ où le terme apparait au moins $k+1$ fois. La lemmatisation rend possible les correspondances quelle que soit la flexion ; la correspondance se fait sur le lemme et non sur sa forme. Par exemple, un terme qui apparaît sous sa forme au pluriel dans la demande de l'utilisateur peut être relié à la forme au singulier renseigné dans le thésaurus. Le recouvrement est donc plus important. Un inconvénient est que cela peut entrainer des rapprochements non satisfaisants, particulièrement dans les cas des uni-termes.

Les *[unknown]* sont les termes qui ne sont pas présents dans les thésaurus. Un *token*

inconnu désigne une entité (ou unité) lexicale qui n'a pas été reconnue lors de l'analyse et les comparaisons avec les termes des thésaurus. Cette information peut soit indiquer que le terme a été mal orthographié faussant l'analyse avec la correspondance des termes issus des thésaurus, soit que le terme n'est pas renseigné dans les thésaurus par manque de précision ou de recouvrement d'un secteur d'activité. Notons que sigles et noms des plus grandes marques ou entreprises peuvent être reconnus si ceux-ci correspondent à des entrées dans le thésaurus. La valeur polysémique, représentée dans le Tableau 7.2, permet d'avoir un aperçu de la polysémie relative à chaque groupe d'utilisateurs selon la tâche de RI.

[POLYSEMY-VALUE]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
0	62,77%	66,60%	78,10%
1	6,82%	7,91%	6,07%
2	2,34%	2,57%	2,11%
3	5,65%	7,11%	3,43%
UNKNOWN	22,42%	15,81%	10,29%

Tableau 7.2 – Valeur polysémique du [REFERENT] de la demande en LN différenciée par type de tâche de RI

Le groupe dont les termes utilisés dans les [REFERENTS] en LN est le moins porteur de polysémie est celui de la [TACHE-PRO] avec un taux de 0 égal à 78,10%. Il a aussi peu de valeurs fortement polysémiques au niveau 3 (3,43%). La [TACHE-SCO] a une valeur polysémique assez élevée en ce qui concerne le niveau 3 avec 7,11%. Les valeurs UNKNOWN sont importantes pour le groupe de la [TACHE-CREA] puisqu'ils sont à 22,42% ; nous ne pouvons donc pas tirer de conclusion quand au lien éventuel entre la longueur des termes (présentée à la Figure 7.7) et leur polysémie.

La complexité linguistique [LINGUISTIC-COMPLEXITY] : la complexité linguistique correspond à la profondeur des nœuds dans le thésaurus du terme ou de l'expression du [REFERENT]. Le [LEVEL 1] correspond aux catégories supérieures généralistes dans la hiérarchie du thésaurus sectoriel : Agro-alimentaire, Biens et Services de Consommation, Industrie lourde, Technologies de l'information et Médias, Sciences de la Vie et Services. Les niveaux suivants de [LEVEL 2] [. . .] [LEVEL 5] sont des niveaux du thésaurus hiérarchiquement descendants. Les termes sont plus spécifiques. Un même terme peut apparaître dans plusieurs branches de la hiérarchie. La distribution des [REFERENTS] par niveaux et par type de tâche de RI est présentée dans le Tableau 7.3 ;

l'usage du thésaurus donne un aperçu de l'utilisation de la profondeur des nœuds et donc de la spécificité de la recherche.

[LINGUISTIC-COMPLEXITY]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
1	2,53%	1,98%	1,06%
2	10,53%	16,40%	9,76%
3	40,74%	35,18%	41,69%
4	19,30%	22,92%	28,76%
5	4,48%	7,71%	8,44%
UNKNOWN	22,42%	15,81%	10,29%

Tableau 7.3 – Complexité linguistique du [REFERENT] de la demande en LN différenciée par type de tâche de RI

D'après le Tableau 7.3, nous pouvons relever que le groupe d'utilisateurs [TACHE-PRO] se distingue particulièrement aux niveaux 4 et 5 ; ils utilisent des termes qui sont plus spécifiques et plus fins que les deux autres groupes. *A contrario*, le groupe [TACHE-SCO] a une demande plus importante au niveau 2 du thésaurus, représentant des domaines plus généraux. Les nombres UNKOWN sont bien sûr identiques à la dernière ligne du Tableau 7.2 de la [POLYSEMY-VALUE] ; il est difficile de dégager des tendances générales pour le groupe [TACHE-CREA]. Ce chiffre élevé de UNKOWN pour la [TACHE-CREA] peut toutefois révéler que les termes employés notamment les entités nommées de noms de marques, d'entreprises ou géographiques sont trop spécifiques (noms de petites entreprises ou noms géographiques d'une petite commune) pour figurer dans les ressources utilisées pour faire les comparaisons.

(c) L'ambiguïté de la tâche : nous avons défini un modèle reprenant certaines valeurs comme la polysémie ou encore la profondeur des nœuds dans le thésaurus afin de rendre compte de l'ambiguïté de la tâche par groupe d'utilisateurs.

Cette ambiguïté se calcule par la somme de fonctions de chacun des traits sémantiques présentés dans la Figure 7.4 : dans la moitié supérieure du tableau, les catégories [ACRONYM], [UNKNOWN], [POLYSEMY VALUE], [FOREIGN] et [STRUCT-SYNTAX-FALSE] font croître la complexité sémantique de la demande en LN.
Dans la moitié inférieure du tableau, les catégories [LINGUISTIC-COMPLEXITY], [CONTEXT-PRECISIONS], [SYNT-DEPTH], [TYPE-DONNEES], [PROPER-NAME-COUNTRY], [NUM-PRICE] et [NUM-DATE] font au contraire décroître cette même complexité : plus elles

sont renseignées moins la demande en LN est ambiguë pour les SRI, c'est-à-dire plus sa compréhension est aisée.

$$ambiguité(tâche) = \sum_{i=1}^{n} f_i(tâche), \tag{7.1}$$

où n est le nombre de traits, f_i est une valeur liée à la tâche et qui pondérera positivement ou négativement un de ses traits. La forme précise de chaque f_i est donnée dans le Tableau 7.5.

trait$_i$	f_i
[ACRONYM]	nombre d'[ACRONYM]
[UNKNOWN]	nombre de [UNKNOWN]
[POLYSEMY VALUE]	valeur de la [POLYSEMY VALUE]
[FOREIGN]	nombre de termes en langue étrangère
[STRUCT-SYNTAX-FALSE]	nombre de demandes avec une structure syntaxique fausse
[LINGUISTIC-COMPLEXITY]	valeur de [LINGUISTIC-COMPLEXITY]
[CONTEXT-PRECISIONS]	nombre de [CONTEXT-PRECISIONS]
[SYNT-DEPTH]	nombre de n-grammes [SYNT-DEPTH]
[TYPE-DONNEES]	nombre de [TYPE-DONNEES]
[PROPER-NAME-COUNTRY]	valeur de la [PROPER-NAME-COUNTRY]
[NUM-PRICE]	nombre de [NUM-PRICE]
[NUM-DATE]	nombre de [NUM-DATE]

Tableau 7.4 – Ambiguité de la tâche de RI en fonction des traits sémantiques

Il en ressort que les groupes [TACHE-SCO] et [TACHE-PRO] ont une valeur quasiment équivalente à 5.7 alors que le groupe [TACHE-CREA] a une valeur a 4.6. Il semblerait donc que l'ambiguïté de la tâche est moins importante pour le groupe [TACHE-CREA]. Ce groupe doit en effet gagner en précision puisque la tâche est souvent assez claire : développer une activité, ouvrir un magasin, etc. La variance est un peu plus importante pour le groupe [TACHE-SCO] ; certains utilisateurs au sein de ce groupe ont une tâche plus ambiguë que d'autres.

A partir de l'estimation de l'ambiguïté de la tâche, nous avons voulu tester si la longueur ([LENGTH] décrite à la page 135) de de la demande en LN était révélatrice de cette ambiguïté. On peut effet émettre l'hypothèse que si une tâche est ambiguë (dans le sens difficile à expliciter voire complexe) l'utilisateur aura tendance à utiliser plus de termes pour l'exprimer. Nous avons réalisé deux mesures pour tester cette hypothèse. Pour la première, nous avons utilisé le coefficient de corrélation de Pearson, indice statistique

[AMBIGUITY-TACHE]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
mean (moyenne)	4.604222	5.758285	5.693676
var (variance)	11.16041	12.41412	11.97133
SD (Ecart-Type)	3.340719	3.523367	3.45996

Tableau 7.5 – Ambiguité de la demande en LN en fonction de la tâche de RI des groupes utilisateurs

qui exprime l'intensité et le sens (positif ou négatif) de la relation linéaire entre deux variables quantitatives. Ce coefficient de corrélation a déjà été utilisé dans d'autres études équivalentes notamment par [MOT 05]. Ces mesures ont été obtenues à l'aide du logiciel *R*. Ce coefficient est nul ($r = 0$) lorsqu'il n'y a pas de relation linéaire entre les variables (ce qui n'exclut pas l'existence d'une relation autre que linéaire). Par ailleurs, le coefficient est de signe positif si la relation est positive (directe, croissante) et de signe négatif si la relation est négative (inverse, décroissante). Ce coefficient varie entre -1 et $+1$; l'intensité de la relation linéaire sera donc d'autant plus forte que la valeur du coefficient est proche de $+1$ ou de -1, et d'autant plus faible qu'elle est proche de 0. L'importance de la valeur de corrélation est exprimée par la valeur *p* associée . $P-valeur$ est une estimation de la probabilité que les résultats aussi extrême ou plus extrême se produisent par hasard . Un $p-valeur$ proche de 0 indique une grande confiance dans la corrélation , tandis qu'une $p-valeur$ proche de 1 indique une forte chance pour l'indépendance entre les variables . Nous retiendrons comme significatives les corrélations dont la $p-valeur$ est inférieure à $0,05$. Les résultats sont donnés dans le Tableau 7.6. Ils indiquent que les [TACHE-SCO] et [TACHE-PRO] ont une $p-valeur$ inférieure à 0,05 et ont donc des résultats significatifs : la longueur de la demande en LN est liée à l'ambiguïté de la tâche. Cette hypothèse est fausse pour la [TACHE-CREA].

	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
Corrélation de Pearson	-0,2067269	-0,2121073	-0,1681577
P-value	0,0501	0,01249	0,0001445

Tableau 7.6 – Ambiguité de la tâche de RI - Coéfficient de corrélation de Pearson

(d) Les secteurs d'activité mentionnées dans les [REFERENTS] : certains secteurs mentionnés dans les [REFERENTS] des demandes en LN sont prédominants selon le type de tâche de RI. Ainsi les utilisateurs du groupe [TACHE-SCO] recherchent davantage

d'informations sur l'alimentation et le textile ; le groupe [TACHE-PRO] sur la construction et BTP ainsi que sur la chimie et les produits chimiques tandis que les utilisateurs de la [TACHE-CREA] recherchent davantage sur les secteurs liés aux loisirs et à la restauration. Ces résultats sont présentés dans la Figure 7.10.

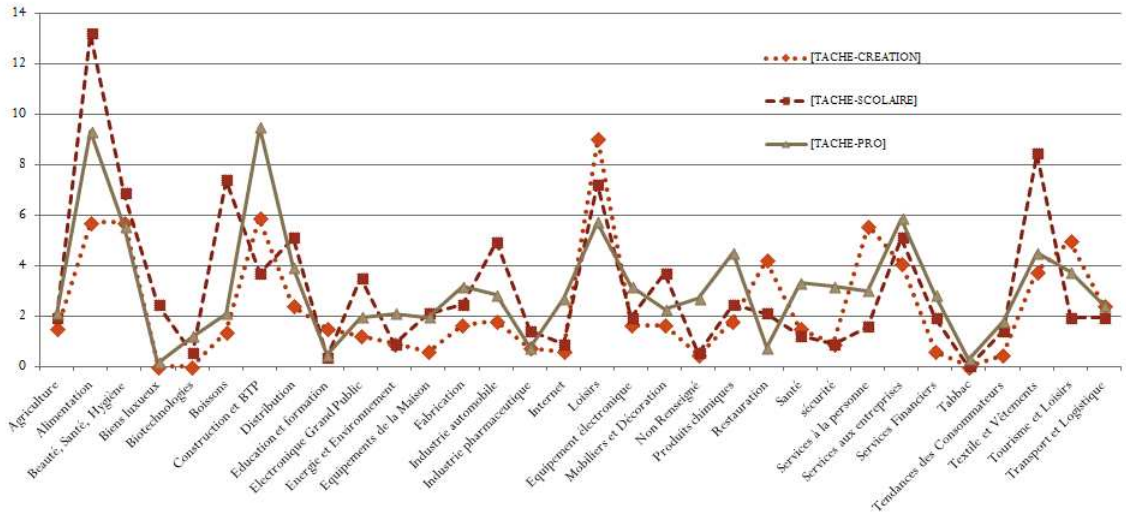

Figure 7.10 – Traits sémantiques des demandes en LN différenciés par type de tâche de RI

7.2 Traits caractéristiques des requêtes selon les types d'utilisateurs

Nous nous intéressons dans cette partie à l'expression des requêtes, différenciée par type de tâches de RI.

7.2.1 Nombre de termes par requête

A la section 6.2.1 (page 109), nous avions noté que la longueur moyenne des requêtes est de 2,4 termes par requête. En analysant les différents profils utilisateurs, nous remarquons que le groupe [TACHE-SCO] a tendance à formuler avec un peu plus de termes (2,9) que les deux autres groupes utilisateurs : [TACHE-CREA] avec 2,1 et [TACHE-PRO] avec 2,2. Soulignons que plusieurs travaux distinguent le nombre de termes par requête en fonction des buts de RI. Par exemple, [MEN 09] indique que le nombre de termes

de la requête varie selon le but de la recherche. les auteurs notent en effet que la majeure partie des requêtes qui possède 5 termes sont des requêtes navigationnelles (*i.e.* dont le but est d'accéder à un site bien particulier et donc dont l'existence est connue ou supposée par l'utilisateur.[1].

Aussi, [PHA 07] étudie le lien entre longueur et degré de spécificité de la requête. Trois expériences sont menées : des participants doivent juger de la spécificité de requêtes formulées par des utilisateurs du moteur de recherche du site Web du gouvernement australien, puis imaginer à leur tour une série de besoins génériques et spécifiques d'informations. Ils formulent ensuite les requêtes correspondantes. Les résultats confirment l'hypothèse que lorsque la longueur de la requête augmente, le besoin d'information de l'utilisateur a tendance à devenir plus spécifique, le compromis entre des requêtes ni trop génériques ni trop spécifiques se situe autour de 3 mots.

Il ressort de ces études un certain consensus entre le trait de longueur de la requête et le besoin d'information de l'utilisateur. Il semble en effet que plus une requête est longue, plus le besoin sous-jacent de l'utilisateur est spécifique et plus la requête est susceptible d'être de type informationnel.

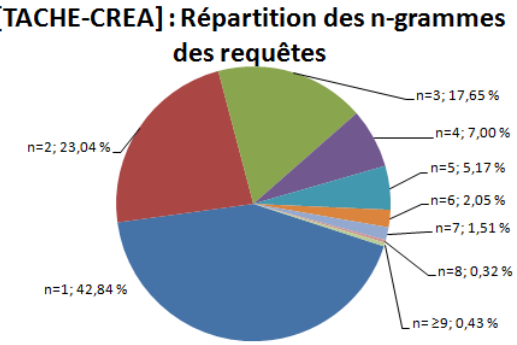

Figure 7.11 – Répartition des n-grammmes dans les requêtes pour le groupe d'utilisateurs [TACHE-CREA]

Les Figures 7.11, 7.12, et 7.13 indiquent la répartition du nombre de n-grammes dans les requêtes par type d'utilisateurs. La répartition est donnée en pourcentage sur l'ensemble des requêtes formulées par les trois types d'utilisateurs en fonction de leur tâche de RI.

[1]selon les trois types types de besoins identifiés par [BRO 02] : Informationel, navigationnel, transactionel, voir à ce sujet le chapitre 1

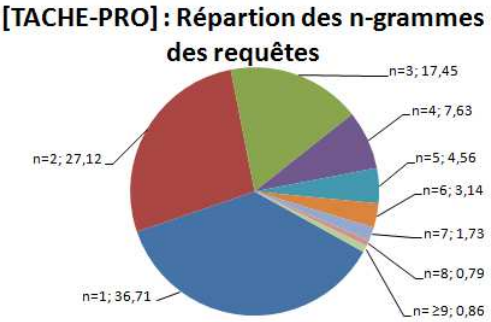

Figure 7.12 – Répartition des n-grammmes dans les requêtes pour le groupe d'utilisateurs [TACHE-PRO]

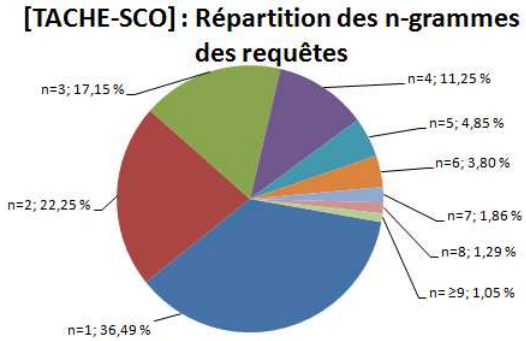

Figure 7.13 – Répartition des n-grammmes dans les requêtes pour le groupe d'utilisateurs [TACHE-SCO]

D'après la Figure 7.11, nous pouvons remarquer que les uni-termes ($n = 1$) sont plus important pour le groupe [TACHE-CREA] à 42,84% que le groupe [TACHE-PRO] à 36,71% ainsi que le groupe [TACHE-SCO] 36,49%. La [TACHE-PRO] a davantage de bi-grammes ($n = 2$) que les autres groupes soit 27,12% puisque les utilisateurs [TACHE-CREA] les utilisent à 23,04% et les [TACHE-SCO] 22,25%. La différence est très faible pour les tri-grammes ($n = 3$) entre les trois groupes d'utilisateurs. Par contre, on peut noter une légère disparité pour les quadri-grammes ($n = 4$) : le groupe [TACHE-SCO] a tendance à légèrement plus utiliser cette structure dans les requêtes : 11,25% alors qu'ils sont de 7,00% pour [TACHE-CREA] et 7,63% pour [TACHE-PRO]. Les différences sur les $n = 5$ et plus sont peu significatives entre les groupes d'utilisateurs.

Pour résumé, le groupe [TACHE-CREA] utilise davantage d'uni-grammes, le groupe [TACHE-PRO] davantage de bi-grammes et le groupe [TACHE-SCO] de quadri-grammes.

En fonction de ces résultats, il peut être intéressant d'établir des pré-supposés sur le profil de l'utilisateur notamment lors d'une première requête ; qui seront ensuite à valider en fonction d'autres critères (prix des études consultés, secteurs des études consultés, etc.).

7.2.2 Traits morpho-syntaxiques des requêtes différenciés par type de tâche de RI

Les résultats sont présentés dans les Figures 7.14 et 7.15. On distingue ainsi quelques préférences d'usage de catégories morpho-syntaxiques.

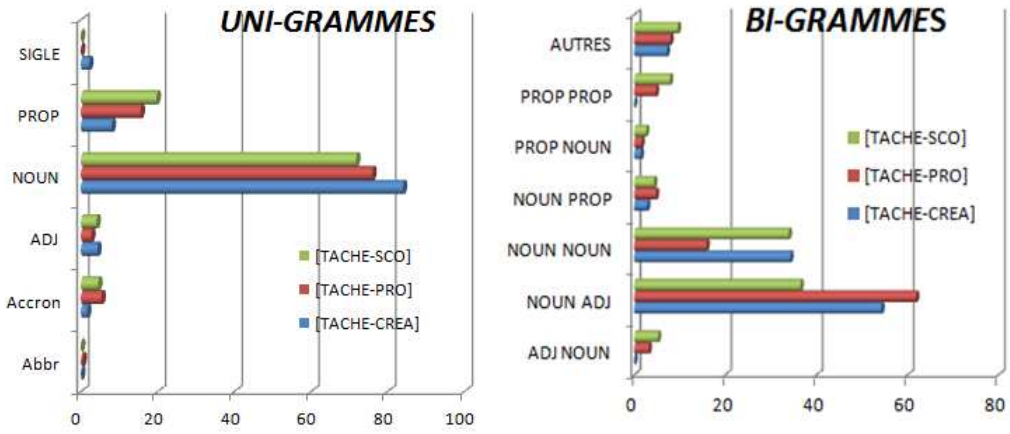

Figure 7.14 – Catégories morpho-syntaxiques différenciées par type de tâche de RI pour les uni- et bi-grammes des requêtes et

D'après la Figure 7.14 et en ce qui concerne les uni-grammes, nous pouvons observer que les trois groupes d'utilisateurs ont une utilisation importante des NOUN (83,96% des uni-grammes pour [TACHE-CREA], 76,00% pour [TACHE-PRO] et 71,75% pour [TACHE-SCO]) et des PROP (19,73% pour [TACHE-SCO], 15,56% pour [TACHE-PRO] et 8,02% pour [TACHE-CREA] quoique de façon moins prononcée. Les autres catégories morpho-syntaxiques sont beaucoup moins représentées. Nous relevons tout de même que le groupe [TACHE-PRO] utilise davantage les [ACCRON] : 5,33% (également important pour le groupe [TACHE-SCO] à 4,48% contre 1,60 pour le groupe [TACHE-CREA]).

Dans les bi-grammes, les NOUN NOUN ainsi que les NOUN ADJ sont également fortement utilisés par les trois groupes. Les NOUN NOUN représentent 34,29% des bi-grammes pour le groupe [TACHE-CREA], 33,91% pour le groupe [TACHE-SCO] contre 15,87% pour le groupe [TACHE-PRO]. Les NOUN ADJ représentent 61,90% des bi-grammes pour les [TACHE-PRO], 54,29% pour les [TACHE-CREA] contre 36,52% pour

les [TACHE-SCO]. A noter donc que le groupe [TACHE-PRO] se distingue *a minima* par l'usage des NOUN NOUN dans les bi-grammes puisqu'ils les utilisent moins que les deux autres groupes tout comme les utilisateurs du groupe [TACHE-SCO] utilisent moins les NOUN ADJ.

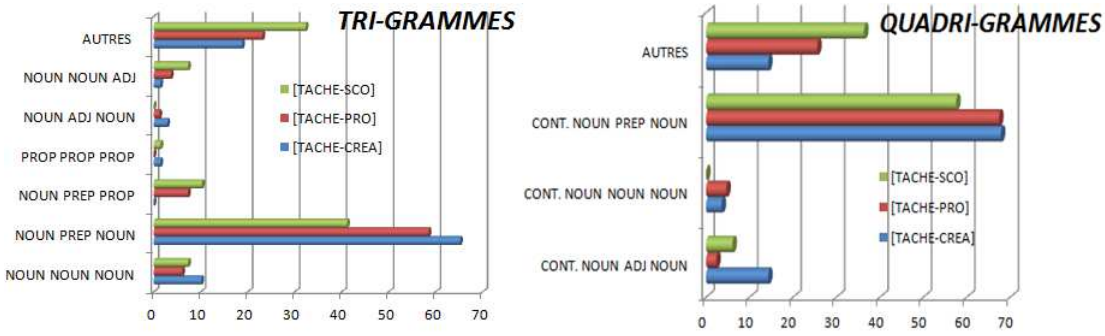

Figure 7.15 – Catégories morpho-syntaxiques différenciées par type de tâche de RI pour les tri- et quadri-grammes des requêtes

Dans les tri-grammes (voir Figures 7.15), les [NOUN PREP NOUN] sont les plus importants : ils représentent 65,22% des tri-grammes des utilisateurs [TACHE-CREA], 58,54% du groupe [TACHE-PRO] et 41,18% du groupe [TACHE-SCO]. On remarque ainsi que les utilisateurs du groupe [TACHE-SCO] utilisent les [NOUN PREP NOUN] mais de façon moins importante que les deux autres groupes.

Également les [NOUN NOUN NOUN] sont bien représentés : 10,14% pour le groupe [TACHE-CREA], 7,35% pour le groupe [TACHE-SCO] et 6,10% pour le groupe [TACHE-PRO]. Nous notons également que les utilisateurs du groupe [TACHE-SCO] utilisent davantage de formes hétérogènes (regroupées dans « Autres ») ainsi que d'autres formes comme des [NOUN NOUN ADJ] et [NOUN PREP PROP].

En ce qui concerne les quadri-grammes, nous notons une forte proportion des [NOUN PREP NOUN] : 67,86% des quadri-grammes pour le groupe [TACHE-CREA], 67,44% pour le groupe [TACHE-PRO] et 57,58% pour la [TACHE-PRO]. Ainsi, les groupes [TACHE-SCO] et [TACHE-PRO] confirment leurs usages des [NOUN PREP NOUN] dans les quadri-grammes ainsi que l'usage de structures plus hétérogènes. De façon parallèle, le groupe [TACHE-CREA] utilise davantage le NOUN ADJ NOUN à la fois dans les tri- et quadri-grammes.

La synthèse des catégories morpho-syntaxique utilisées dans les requêtes et différenciées par groupes d'utilisateurs en fonction de leur tâche de RI est présentée dans le

Tableau 7.16.

REQUETE	Uni-grammes	Bi-grammes	Tri-grammes	Quadri-grammes
[TACHE-CREA]	NOUN PROP	NOUN NOUN NOUN ADJ	NOUN PREP NOUN NOUN NOUN NOUN	NOUN PREP NOUN +1 NOUN ADJ NOUN +1
[TACHE-SCO]	PROP ADJ	NOUN NOUN PROP PROP	NOUN PREP PROP NOUN NOUN NOUN	NOUN PREP NOUN +1
[TACHE-PRO]	ACCRON NOUN	NOUN ADJ	NOUN PREP NOUN NOUN PREP PROP	NOUN PREP NOUN +1

Figure 7.16 – Synthèse des catégories morpho-syntaxiques des requêtes et différenciés par type de tâche de RI

7.2.3 Présence d'informations spécifiques dans la requête selon la tâche de RI

Nous avons comparé dans cette sous-section les spécificités des équations de recherche selon la tâche de RI. Ces spécificités sont présentées dans la Figure 7.17 : l'usage des opérateurs booléens dans les requêtes [BOOLEAN-OPERATORS], les requêtes en langue étrangère [FOREIGN-LANGAGE], l'indication de noms de pays (précisions géographiques) [PROPER-NAME-COUNTRY], l'indication de noms d'entreprises ou de marques [PROPER-NAME-BUS], une indication de date [NUM-DATE], une indication numérique (quantité, volume, référence...) [NUM-QUANTITY] et enfin le type de données recherché (swot, fusion et acquisition, statistiques, chiffres d'affaires...) [DATA-TYPE].

D'après la Figure 7.17, nous observons que les utilisateurs du groupe [TACHE-PRO] utilisent de façon plus prononcée certaines spécificités dans les équations de recherche ; à savoir : les noms propres de type géographique (PROPER-NAME-COUNTRY), les noms propres de types Business (PROPER-NAME-BUS). Ils indiquent aussi plus fortement les types de données attendues (DATA-TYPE). Également de façon moins prononcée, un usage à peine plus développée des opérateurs booléens que les deux autres groupes d'utilisateurs.

Pour les opérateurs booléens (BOOLEAN-OPERATORS) et pour l'usage de la langue étrangère (FOREIGN-LANGAGE), il n'est pas possible d'effectuer des comparaisons avec

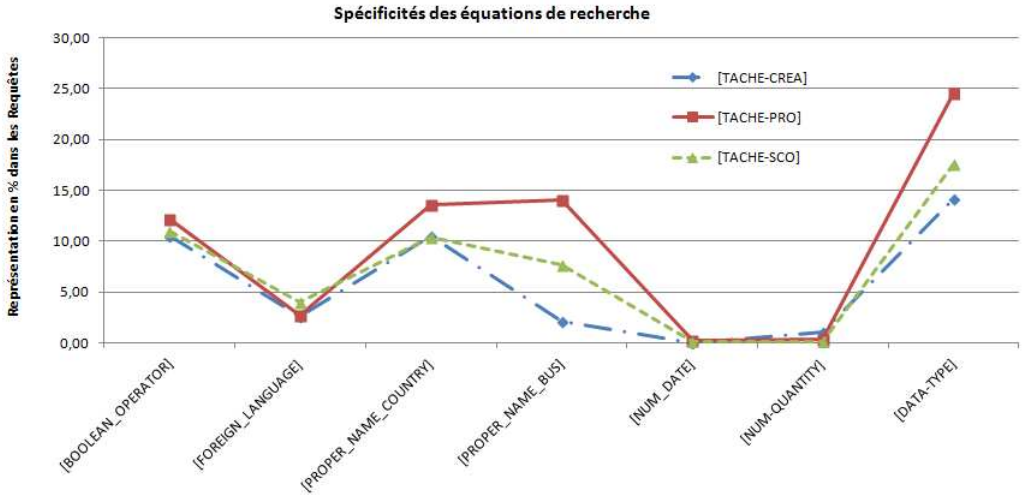

Figure 7.17 – Spécificités des équations de recherche différenciées par type de tâche de RI

le [REFERENT] car ce sont des spécificités qui étaient alors inexistantes. *A contrario*, nous pouvons comparer les PROPER-NAME-COUNTRY, PROPER-NAME-BUS, NUM-DATE, NUM-QUANTITY et DATA-TYPE. Il apparait que si dans les requêtes, ce sont les utilisateurs ayant une [TACHE-PRO] qui utilisent davantage les PROPER-NAME-COUNTRY, PROPER-NAME-BUS et les DATA-TYPE, il n'en est pas de même pour le [REFERENT]. Le groupe ayant une [TACHE-CREA] utilise davantage les PROPER-NAME-COUNTRY et les utilisateurs ayant une [TACHE-SCO] emploient davantage les PROPER-NAME-BUS.

Nous remarquons certaines pertes d'informations (car moins fréquentes) dans les requêtes :

- les PROPER-NAME-COUNTRY *i.e.* tout ce qui est indications géographiques pour le groupe [TACHE-CREA],
- les PROPER-NAME-BUS *i.e.* les noms de marque ou d'entreprises pour le groupe [TACHE-SCO],
- les NUM-PRICE alors très présent dans les [REFERENTS] notamment par les utilisateurs ayant une [TACHE-SCO] sont complètement inexistants dans les requêtes.

Les utilisateurs ayant une [TACHE-PRO] sont ceux qui ajoutent le plus d'informations dans les requêtes. Ces derniers exploitent mieux les fonctionnalités de recherche notamment *via* l'usage des opérateurs booléens mais aussi *via* la contextualisation de

leur besoin informationnel. C'est le groupe d'utilisateurs qui ressemble le plus à un usage "expert" (dans le sens compétence et maitrise des techniques de RI) des moteurs de recherche.

7.3 Conclusion : Comparaison de la demande en LN à la requête avec différenciation par type de tâches de RI

Nous reprenons dans cette section :

1. les correspondances (totales ou partielles) entre le [REFERENT] et la requête,
2. les correspondances (totales ou partielles) entre les autres blocs de la demande en LN et la requête,
3. les correspondances *via* des relations sémantiques entre les [REFERENTS] et les requêtes détectées par des ressources lexicales externes,
4. l'absence de correspondances entre les [REFERENTS] et les requêtes

Nous présentons les résultats dans le Tableau 7.7. Les correspondances totales entre les [REFERENTS] et les requêtes sont assez importantes quel que soit le groupe d'utilisateurs.

Les Figures 7.8 et 7.9 présentent en détail les correspondances partielles des demandes et des relations sémantiques détectées différenciées par type de tâche de RI.

Les correspondances partielles sont moins importantes pour le groupe [TACHE-SCO] avec seulement 24,17% de correspondances partielles entre le [REFERENT] et la requête. En regroupant les correspondances totales et partielles entre le [REFERENT] et la requête, nous arrivons à un taux très important de recouvrement *i.e.* où le nombre de termes en commun entre ces deux ressources d'information sont de 68,34% pour la [TACHE-CREA], 61,21% pour la [TACHE-SCO] et 74,31% pour la [TACHE-PRO].

Plus précisément sur les correspondances partielles (Tableau 7.8), nous observons que très peu d'utilisateurs du groupe [TACHE-SCO] font des variations morphologiques entre le REFERENT et la requête alors que ces variations sont plus importantes pour les utilisateurs du groupe [TACHE-PRO] notamment les transformations flexionnelles. Les utilisateurs du groupe [TACHE-SCO] semblent effectuer peu de transformations entre les deux types d'énoncés, gardant dans 44,84% des cas les termes identiques entre la demande en LN (toutes correspondances totales) contre 40,11% pour les utilisateurs de

	[TACHE-CREA] (sur 379)	[TACHE-SCO] (sur 513)	[TACHE-PRO] (sur 506)
Correspondances totales [REFERENT] et requête	132 (34,83%)	190 (37,04%)	181 (35,77%)
Correspondances partielles [REFERENT] et requête	127 (33,51%)	124 (24,17%)	195 (38,54%)
Correspondances totales autres blocs et requête	20 (5,28 %)	40 (7,80%)	11 (2,17%)
Correspondances partielles autres blocs et requête	35 (9,23 %)	78 (15,20%)	11 (2,17%)
Relations sémantiques détectées via autres ressources	55 (14,51%)	69 (13,45%)	89 (17,59%)
Environnement sémantique inexistant	10 (2,64%)	12 (2,34%)	19 (3,75%)

Tableau 7.7 – Correspondances entre les demandes en LN et les requêtes différenciées par type de tâche de RI

la [TACHE-CREA] et 37,94% pour les utilisateurs de la [TACHE-PRO]. Les utilisateurs ayant une [TACHE-PRO] et [TACHE-SCO] opèrent davantage de changements notamment de synonymie, d'hyperonymie et d'hyponymie (Figure 7.9).

Les correspondances (partielles ou totales) entre les autres blocs de la demande en LN et la requête sont de 14,51% pour la [TACHE-CREA], 23,00% pour la [TACHE-SCO] et 4,34% pour la [TACHE-PRO]. Ce dernier chiffre est à mettre en correspondance avec le fait que les utilisateurs ont peu recours à ces blocs (voir aussi la Figure 7.3, page 137). Ainsi les utilisateurs ayant une tâche professionnelle utilisent très peu les autres blocs pour mentionner le ou les terme(s) de leur recherche. Par contre, cet usage est assez important pour les utilisateurs ayant une tâche scolaire ; il apparaît donc intéressant pour un SRI de récupérer ce genre d'informations en proposant par exemple une aide personnalisée de (re)formulation des requêtes.

Détails Correspondances partielles	[TACHE-CREA] (sur 379)	[TACHE-SCO] (sur 513)	[TACHE-PRO] (sur 506)
[A] Variations morphologiques	**36 (9,50%)**	**25 (4,87%)**	**70 (13,83%)**
[A.1] Transformations flexionnelles	19 (5,01%)	15 (2,92%)	49 (9,68%)
[A.2] Transformations dérivationnelles	17 (4,49%)	10 (1,95%)	21 (4,15%)
[B] Modifications de la langue	**0 (0 %)**	**5 (0,97%)**	**11 (2,17%)**
[C] Modifications d'un ou plusieurs terme(s) dans le [REFERENT]	**91 (24,01%)**	**94 (18,32%)**	**114 (22,53%)**
[C.1] Ajouts de termes	14 (3,69%)	26 (5,07%)	15 (2,96%)
[C.2] Glissements de termes	16 (4,22%)	10 (1,95%)	17 (3,36%)
[C.3] Suppressions de termes	61 (16,09%)	58 (11,31%)	82 (16,21%)

Tableau 7.8 – Détails des correspondances partielles des demandes en LN et des requêtes différenciées par type de tâche de RI

Détails Relations Sémantiques	[TACHE-CREA] (sur 379)	[TACHE-SCO] (sur 513)	[TACHE-PRO] (sur 506)
Synonymie	37 (9,76%)	35 (6,82%)	47 (9,29%)
Hyperonymie	13 (3,43%)	9 (1,75%)	19 (3,75%)
Hyponymie	3 (0,79%)	4 (0,78%)	6 (1,19%)
Métonymie	2 (0 ,53%)	21 (4,09%)	17 (3,36%)

Tableau 7.9 – Détails des relations sémantiques détectées différenciées par type de tâche de RI

7.4 Conclusions générales et perspectives

Le processus mis en place nous a permis de segmenter tout d'abord les demandes en LN issues des formulaires de SAV en blocs d'informations puis de les transformer

en concepts *via* la cartouche de connaissance de Luxid. L'analyse effectuée a relevé les spécificités des demandes en LN différenciées par le type de tâche nous permettant d'obtenir à un instant *T* une analyse de l'expression des besoins utilisateurs.

Les demandes en LN ainsi transformées ont été comparées aux requêtes effectuées dans le moteur de recherche *Plusdetudes.com*. Les résultats obtenus sont les suivants :

- Environ 83% des termes utilisés pour décrire la thématique sont les mêmes dans les demandes en LN et dans les requêtes (Hypothèse [H1]) (la moitié avec une correspondance partielle) dont l'ordre reste inchangé (Hypothèse [H2]) (Figure 7.10),
- Environ 15% des requêtes présentent un environnement sémantique proche,
- Environ 3% ne présentent ni correspondance ni environnement sémantique (Figure 7.11).
- Les demandes en LN conservent des indications de la tâche de RI (Hypothèse [H3]) : thématique + géo + type de données + noms de marque/société
- Par contre, les requêtes conservent peu d'indications de la tâche de RI (il y a donc une déperdition d'informations dans les requêtes). Plus précisément :les utilisateurs ayant une [TACHE-SCO] ne mentionnent plus la contrainte de budget (contrainte présente dans les demandes en LN), les utilisateurs ayant une [TACHE-CREA] mentionnent moins la zone géographique dans les requêtes que dans les demandes en LN. A contrario, les utilisateurs ayant une [TACHE-PRO] mentionnent **davantage** les noms de marque/société (les requêtes s'avèrent alors être plus longues).

Nos résultats montrent la nécessité de récupérer des informations comme les types d' informations recherchées, les objectifs visés, les contraintes associés, etc pour les réinjecter dans les requêtes utilisateurs (et ceci plus spécifiquement sur certains profils comme les utilisateurs ayant une [TACHE-SCO] et [TACHE-CREA]). L'obtention des contextes (utilisateur et tâche de recherche d'informations) s'avère donc primordiale pour les SRI qui doivent mettre en place un système plus efficace d'interrogation. Plusieurs pistes sont envisageables comme construire un profil utilisateur d'après les requêtes obtenues et adapter le SRI en fonction (un exemple d'application est proposé à la Figure 7.12, proposer à l'utilisateur un système hybride entre la LN et les requêtes ou bien un système d'agents conversationnels appliqués sur des requêtes ou encore un

	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
Correspondances totales	40 %	45 %	38 %
Correspondances partielles	43 %	39 %	41 %
TOTAL	83 %	84 %	79 %

Tableau 7.10 – Conclusions obtenues sur la comparaison des énoncés : formulaire SAV *versus* requêtes

	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
Avec Environnement Sémantique	14%	14 %	18 %
Sans Environnement Sémantique	3 %	2 %	4 %

Tableau 7.11 – Conclusions obtenues sur la comparaison des énoncés : avec environnement sémantique *versus* sans environnement sémantique

système de facettes et d'applications de filtres au fur et à mesure des recherches des utilisateurs.

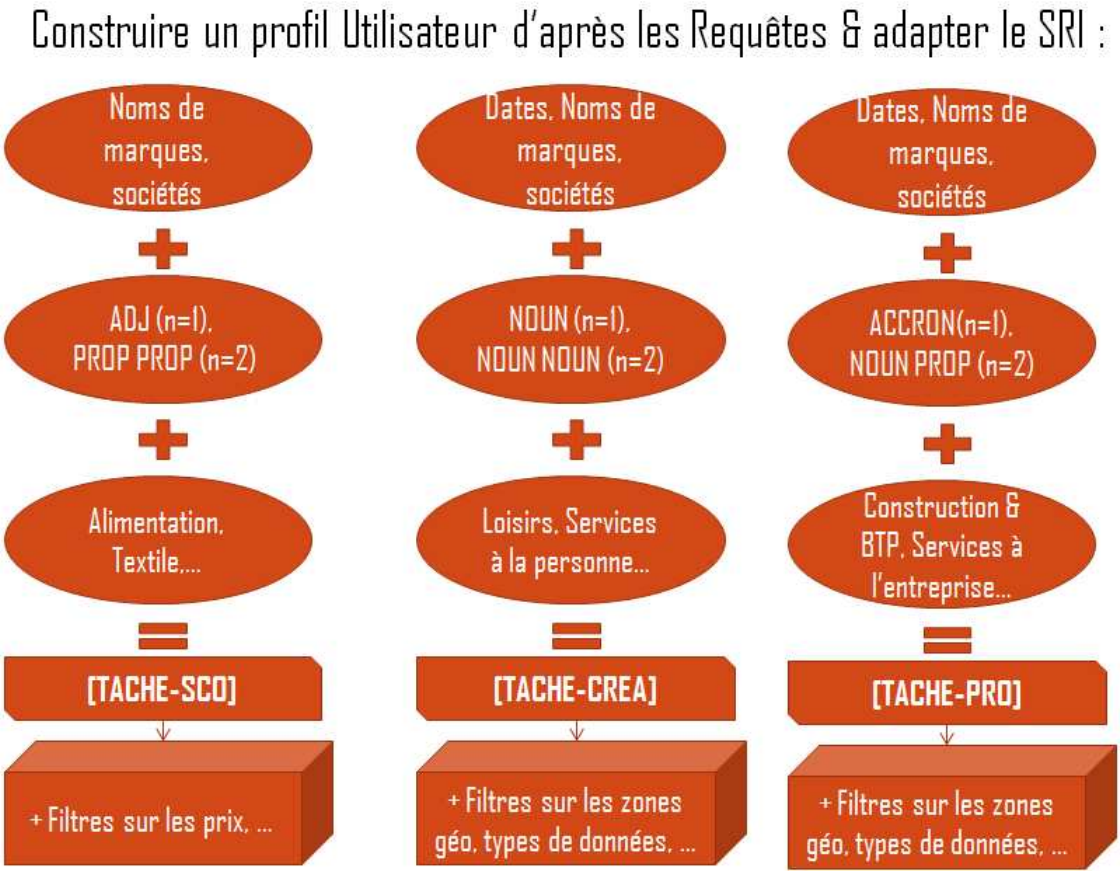

Tableau 7.12 – Perspectives : Construire un profil utilisateur d'après les Requêtes et adapter le SRI

CONCLUSION

Le besoin informationnel d'un utilisateur se formule soit en LN (correspondant ici à un langage d'« expression libre » et « sans contrainte ») soit en langage de requêtes (sous forme de mots-clés *via* un formulaire de recherche d'un SRI). Les mécanismes de sa verbalisation sont eux mêmes régis par certaines règles liées au contenu informatif recherché. La question principale qui a motivé le développement des méthodes présentées dans cette thèse a été précisément centrée sur l'étude des relations entre ces mécanismes. L'objectif de notre approche étant d'évaluer la façon dont un même utilisateur verbalise un besoin informationnel à travers un énoncé de type langage naturel et un énoncé de type langage de requêtes.

Une première approximation serait de penser que l'expression en LN est une verbalisation plus riche que le besoin informationnel à proprement parler. Nous avons développé des méthodes permettant de déterminer quelles informations sont réellement ajoutées, modifiées, supprimées entre la demande en LN et la requête.

La littérature actuelle sur les typologies de requêtes, l'analyse linguistique des requêtes ou les dialogues homme-machine est très fournie. En revanche, il n'existe pas, à notre connaissance, de travaux comparant les deux types d'énoncés précédemment cités. Ceci s'explique bien sûr par le fait que seulement dans de rares occasions l'utilisateur est appelé à formuler son besoin informationnel par ces deux types d'énoncés. Malgré le manque de références, de méthodes pour comparer ces deux formes d'énoncés, les résultats que nous avons obtenus permettent de clarifier certains aspects : les besoins informationnels et la façon dont les utilisateurs les expriment à travers ces deux types d'énoncés sont intimement liées, ce que nous mettons en évidence dans notre travail sous un jour nouveau.

Résultats

Dans le chapitre 4, nous avons recueilli et analysé les besoins informationnels exprimés à la fois en LN et en langage de requêtes par les mêmes utilisateurs, ceci dans un contexte bien particulier ; celui d'une demande de remboursement effectuée par des utilisateurs d'un moteur de recherche après avoir exécuté des requêtes *a priori* infructueuses. Pour cela, nous nous situons dans le contexte applicatif d'un moteur de recherche dédié à des études économiques en français. Nous avons recueilli *via* ce moteur de recherche, tous les besoins informationnels exprimés à la fois en langage naturel et en langage de

requêtes sur 5 années consécutives (de 2002 à 2007). Nous avons totalisé un corpus de 1398 demandes en langue naturelle et de 3427 requêtes (une demande en langue naturelle pouvant être formulée par une ou plusieurs requêtes de la part de l'utilisateur).

Nous nous sommes attaqués dans le chapitre 5 aux règles linguistiques et une analyse morpho-syntaxique qui nous a permis de schématiser l'énoncé en LN, relever des informations de contexte et enfin d'obtenir une représentation sémantiquement comparable aux requêtes. Nous avons proposé une méthode entièrement automatique. Celle-ci se décompose en trois étapes :

(1) une phase de segmentation des demandes en LN en blocs d'informations : nous avons établi que les demandes en LN formulées dans les formulaires SAV comportent pour la plupart une structure sous-jacente, s'articulant autour de blocs d'informations (voir page 70). L'hypothèse fondamentale est qu'un analyseur peut pré-traiter le travail de repérage de ces blocs pour un domaine bien particulier en repérant et en analysant les régularités observées dans notre corpus pour effectuer des découpages et des structurations.

Certaines formulations du besoin informationnel ne nécessitent pas l'ensemble de ces blocs. Des blocs peuvent également se scinder : nous avons pu ainsi relevé trois emplacements possibles pour le bloc [CONTEXT]. Une structure alternative de demande est également présente : celle interrogative pour laquelle un autre schéma d'analyse organisé également en blocs d'informations est proposé.

Pour effectuer la segmentation en blocs d'informations, nous avons dégagé à partir d'un corpus d'apprentissage les régularités des structures énonciatives afin de mettre au point des procédures qui s'appuient sur les principaux motifs séquentiels fréquents. Dans un premier temps, notre méthodologie a consisté à extraire de manière empirique les caractéristiques et les structurations des *items* d'un corpus d'apprentissage constitué d'environ 15 % de notre corpus total soit 200 demandes en LN. Pour cela, nous avons relevé manuellement le vocabulaire et les formulations d'une demande pour les différents blocs d'informations en nous appuyant sur des marqueurs sémantiques ainsi que sur des marqueurs typographiques comme la ponctuation et les majuscules. Dans un second temps, nous avons procédé à l'automatisation de cette segmentation. Pour notre étude, cette automatisation a été implémentée grâce à la technologie *Skill Cartridge*^{TM}(également appelé *Cartouches de Connaissances*) de l'outil *Luxid*. C'est d'une solution industrielle que nous avons adapté à notre propre corpus. Cette solution a été conçue pour rechercher et extraire des informations dans des données non structurées à partir de règles

linguistiques et de lexique. Les règles linguistiques permettent d'ordonner l'apparition des différents blocs d'informations qui sont transformés *via* les *Cartouches de Connaissances* en *concepts*. L'évaluation manuelle de cette méthode a montré que dans 75 % des cas le découpage en motifs obtenu était correcte ; 20 % des séquences restaient non étiquetées (*i.e.* pas d'appartenance aux différents blocs d'informations identifiés) et 5 % des séquences étaient étiquetées de manière erronée. Si l'on effectue un tri supplémentaire, manuel cette fois, on peut distinguer, plusieurs cas de figures : *(1)* des mauvais étiquetages, principalement sur les noms propres (problèmes de majuscules), *(2)* des mauvaises structurations de l'utilisateur probablement dues à des erreurs lors des phases de corrections orthographiques et typographiques de certains termes.

Les séquences étiquetées de manière erronée ou non étiquetées ont été réaffectées manuellement.

L'avantage de cette approche est qu'elle ne nécessite ni corpus annoté manuellement ni analyse morpho-syntaxique. Elle nous a permis *(1)* de conserver l'ordre des énoncés de la demande, *(2)* de travailler transversalement sur certaines parties du discours (comme les référents), *(3)* d'automatiser le processus afin de prévenir d'éventuels erreurs de catégorisation manuelle lors de la répartition des données dans les différents concepts (même si ces erreurs peuvent être peu nombreuses), *(4)* de mettre en place une méthode qui puisse être exploitée assez rapidement sur d'autres corpus et donc à plus grande échelle. Néanmoins, cette méthode nécessite une sélection des motifs pertinents car le nombre de motifs extraits peut être assez important surtout en fonction du corpus test. En parallèle de ce travail de segmentation, nous avons également mis en place une méthode d'extraction d'entités nommées (EN) quelle que soit sa place dans la demande en LN à savoir : la zone géographique [PROPER-NAME-COUNTRY], le nom des entreprises et/ou de marques [PROPER-NAME-BUS], les expressions de temps [NUM-DATE], de quantités, pourcentages [NUM-QUANTITY], les valeurs monétaires [NUM-PRICE]. Les EN sont alors désignées comme des éléments atomiques dans le texte appartenant aux catégories énumérées ci-dessus. Pour extraire ces EN, la technologie des *cartouches de connaissances* combine à la fois le recours à du vocabulaire sous forme de lexique et à la fois à des règles linguistiques qu'il faut programmer en fonction du contexte.

(2) une utilisation différenciée des blocs d'informations de la demande ne LN : nous avons analysé les différents blocs d'informations en deux temps :

1. une analyse linguistique de certains traits morphologiques, syntaxiques et séman-

tiques des blocs d'informations de la demande en LN. Ces traits permettent d'obtenir des informations de type « général » sur la demande elle-même comme sa longueur, son complexité, sa structure grammaticale ou encore le nombre et le type d'entités nommées,

2. une analyse syntaxique plus fine du bloc [REFERENT], considéré dans cette étude comme le pendant de la requête. Cette analyse nous permet d'avoir une vision plus précise sur la composition du [REFERENT] comme le nombre et la structures des termes.

Rappelons que nous avons effectué cette même analyse syntaxique sur les requêtes afin de pouvoir établir des comparaisons.

(3) une comparaison des demandes en LN et des requêtes : en effet, un atout majeur de notre étude a été de pouvoir comparer les requêtes et les [REFERENTS] de la demande en LN identifiés lors de la phase de segmentation en blocs d'informations. Nous avons alors pu relever des régularités le cas échéant, entre les deux types d'expression du besoin informationnel. Nous avons repris les travaux de [HUA 09] pour concevoir manuellement une grammaire afin de générer automatiquement la comparaison des formes de surface (pas d'interprétation sémantique) des [REFERENTS] et des requêtes.

Notre hypothèse H1 qui consistait à valider si le choix du vocabulaire fait pour la saisie de la requête correspond aux termes employés lors de la formulation de la demande en LN a été validée.

En effet, nos résultats indiquent peu de différences sur le vocabulaire utilisé entre les [REFERENTS] et les requêtes ; ils consistent souvent en la répétition et la réutilisation des mêmes termes avec toutefois : *(i)* l'augmentation ou la diminution de quelques termes, *(ii)* la transformation (flexionnelle, plus rarement dérivationnelle) de quelques termes, *(iii)* l'utilisation de formes dérivées ou d'abréviations, *(iv)* l'association de quelques synonymes ou concepts voisins par rapport au domaine visé par la requête initiale.

L'hypothèse H2 qui consistait à tester si l'ordre des termes de la question en LN est identique à la structure de la requête a elle aussi été validée puisque très peu d'utilisateurs modifient l'ordre d'apparition des termes ou des expressions entre le [REFERENT] de la demande en LN et la requête. Seulement 24 requêtes (soit 1,71%) indiquent un inversement de l'ordre entre les mots de la demande en LN et les termes de la requête ; 8 sont dus à un passage d'une forme adjectivale au nom propre du pays (ex : « café en

France » => « marché français du café » ou encore « poisson surgelé en Espagne » => « marché espagnol du poisson surgelé ».

Il n'existe apparemment pas d'apport majeur à utiliser les demandes en LN à proprement parler sur la formulation du secteur d'activité. En revanche, nous observons un apport indéniable dans la demande en LN dans les blocs [CONTEXT] et [PRECISIONS] qui permettent d'obtenir plusieurs informations complémentaires très intéressantes : la zone géographique, le type de données économiques désirés, le budget. Bien que connexes, ces informations sont riches pour contextualiser le besoin informationnel et plus spécifiquement pour déterminer les critères de sélection d'études de marché. Ce sont des informations qui doivent être récupérées soit avec une aide à la formulation des besoins (type agents conversationnels) soit à partir du profil enregistré par l'utilisateur.

Nous avons finalement clos la présentation des résultats avec le chapitre 7 où nous mettons en lumière les résultats sus-cités sous l'angle des tâches de RI à accomplir.

Notre hypothèse H3 qui consistait à tester si la demande en LN ainsi que la requête conservaient des éléments inhérents à la tâche à effectuer par l'utilisateur a été partiellement validée.

Dans la première partie de notre hypothèse H3, l'hypothèse était que les demandes en LN comportent des indices linguistiques et structurels permettant de repérer les buts explicites des tâches de RI des utilisateurs ; indices qui se seraient également présents dans la requête. Nous avons en effet plus spécifiquement comparé les deux types d'énoncés en fonction de trois classes d'utilisateurs : [TACHE-CREA] dont le but de la RI est la création d'entreprise ou le lancement d'un nouveau produit ou encore d'une nouvelle marque, [TACHE-SCO] dont le but de la RI est la réalisation d'une tâche scolaire ; [TACHE-PRO] dont le leur but de la tâche de RI est l'obtention d'informations dans un cadre professionnel. Or, grâce à notre corpus très spécifique, nous avons en effet observé des régularités sur la construction des demandes en LN : à la fois dans les structures mais aussi dans la nature même des informations.

Dans la seconde partie de notre hypothèse H3, nous voulions tester si les indices linguistiques et structurels permettant de repérer les buts explicites des tâches de RI étaient également présents dans les requêtes. Les résultats n'ont validés que partiellement la deuxième partie de cette hypothèse. A part quelques exceptions, nous observons dans les structures linguistiques des requêtes seulement des différences mineures entre les trois groupes. En effet, ces trois groupes utilisent en majorité des NOUN dans les unigrammes et des NOUN NOUN dans les bi-grammes, des NOUN PREP NOUN dans les tri et quadri-grammes.

Les informations relevées dans les demandes en LN ne sont plus que partiellement présentes dans les requêtes : elles le sont pour les utilisateurs [TACHE-PRO] qui présentent des requêtes assez renseignées (voir plus) que dans les demandes en LN ; ils maîtrisent par ailleurs mieux les fonctionnalités de recherche du moteur de recherche. En revanche, les utilisateurs [TACHE-CREA] mentionnent moins souvent les zones géographiques dans les requêtes que dans les demandes en LN. De même, les utilisateurs [TACHE-SCO] mentionnent moins souvent les noms de marque ou d'entreprises dans les requêtes ; ces derniers ne mentionnent pratiquement pas non plus les critères de prix dans les requêtes alors que cette information était très présente dans les demandes en LN et est un critère important dans la sélection d'études de marché. Nous ne retrouvons donc pas d'indices linguistiques et structurels des buts explicites des tâches de RI dans les requêtes puisque les profils identifiés adaptent des comportements différenciés : les utilisateurs [TACHE-PRO] ont tendance à être plus précis et plus explicites que dans les demandes en LN, de mieux exploiter les fonctionnalités de recherche par rapport aux deux autres groupes utilisateurs. Les utilisateurs [TACHE-CREA] et [TACHE-SCO] mentionnent beaucoup plus rarement dans les requêtes les critères importants issus de la demande en LN et spécifiques à leurs tâches de RI (zone géographique, prix, délais, noms de marque ou d'entreprises bien spécifiques).

Discussion

Nous nous sommes confrontés à plusieurs problèmes -notamment techniques- tout au long de cette thèse et nous avons fait certains choix pour les résoudre. D'autres approches auraient été possibles. Tout d'abord, notre thèse se situe au carrefour de plusieurs disciplines : modélisation utilisateur, recherche d'informations, interaction homme-machine, informatique, traitement automatique de la langue (TAL), cognition ou encore linguistique. Multiplier le recours à des disciplines variées, c'est s'approprier leurs méthodologies. Plus précisément, notre travail, d'abord ancré en RI a fait beaucoup appel au TAL et à la linguistique, ce qui fait sa richesse mais aussi probablement sa complexité. Pour aller plus loin, nous aurions pu étudier les régularités lexicales, syntaxiques et énonciatives des énoncés, rendre compte des phénomènes de dérivation ou encore utiliser des arbres et/ou des descriptions d'arbres, des analyseurs syntaxiques robustes en dépendance pour aller plus loin dans les analyses linguistiques. Mais notre objectif était tout d'abord de concevoir un outil « réaliste » et pas trop gourmand en terme de temps de traitement. Par ailleurs, l'ampleur des analyses syntaxiques à réaliser aurait largement dépassé le

cadre de nos compétences pour ce travail. Cela peut éventuellement être une piste pour d'éventuels prolongements de nos travaux.

Dans le même sens, même si notre découpage en blocs d'informations est assez efficace, nous avons hésité à étudier de façon systématique et plus complète la totalité de la demande en LN. Notre approche était motivée par le fait que toute la demande n'était pas nécessairement intéressante à analyser (multiplicité d'exemples, longueur des formulations de politesse, etc.), mais aussi demandait une partie importante et assez lourd de traitements linguistiques. Notre objectif était d'automatiser le schéma de segmentation des demandes en LN afin de savoir où aller chercher les informations nécessaires pour compléter/contextualiser un besoin informationnel. Même si dans un dernier temps notre approche avait pour objectif d'obtenir ces résultats, une amélioration possible serait donc d'étudier avec d'autres processus la totalité de la demande en LN pour observer si d'autres éléments intéressants à prendre en compte pour le SRI émaneraient.

Nous nous sommes ensuite focalisés dans ce travail sur les expressions du besoin informationnel à travers la demande en LN et la (les) requête(s) ; nous pensons que la prise en compte d'autres variables comme les documents consultés, le nombre de clics, la durée de consultation sont des métriques aidant à la compréhension et à l'amélioration de la construction du profil utilisateur. L'étude de ces éléments nécessiterait un travail complémentaire certainement très intéressant et instructif qui viendrait compléter les résultats déjà obtenus dans cette thèse.

Enfin, notre travail s'est effectué sur un corpus en français ; beaucoup de travaux portant les (re)formulations de requêtes sont en anglais. Afin de pouvoir comparer plus facilement notre méthode et les résultats obtenus avec d'autres travaux existants, nous envisageons de déployer notre méthode sur le service SAV de l'entreprise Ubiquick (moteur ReportLinker) mais en anglais cette fois. Il serait alors intéressant de comparer les différences d'expression entre plusieurs langues mais c'est un tout autre travail !

Perspectives

On pourrait imaginer que les SRI puissent construire le profil utilisateur en fonction des éléments que nous avons utilisé : secteur d'activité, types d'informations recherchées, nombre de termes dans les requêtes, ainsi qu'en fonction de certains traits caractéristiques des [REFERENTS] comme les catégories morpho-syntaxiques des n-grammes. A partir de certains indicateurs identifiés ci-dessus ainsi que d'autres métriques que nous n'avons pas étudié ici comme le nombre de documents consultés, la durée de consulta-

tion,etc.), les SRI pourraient alors affiner la construction du profil afin de construire une architecte de recherche personnalisée et adaptée aux besoins et compétences utilisateurs. L'intérêt de cette approche est de construire une approche par facettes (de prix, de zone géographique, de types de données...) qui soient activées et affichées à l'utilisateur qu'en fonction de ce que le SRI ait pu déduire de ses actions. [ING 05] et [LI 08] proposent des modèles à facettes qui aspirent à faciliter la recherche.

D'après nos conclusions, toutes les requêtes ne sont pas à traiter de façon identiques par les SRI : à l'instar des travaux de [STR 08], certaines requêtes sont suffisamment explicites sans compléments, notamment celles qui comportent plus de deux termes. Les enjeux et les perspectives seraient alors non pas de proposer une aide systématique de reformulation ou d'aide à la formulation mais bien d'aider à la conceptualisation du besoin informationnel en fonction du profil et de la tâche à effectuer. Pour certains profils (nos [TACHE-SCO] et [TACHE-CREA]), pour des requêtes trop courtes (moins de deux termes) ou encore lorsque le besoin informationnel est encore vague, il est intéressant de garder une formulation plus libre (ou autre que les requêtes) pour obtenir un maximum d'informations et aider à la contextualisation du besoin informationnel. Il est nécessaire dans ces cas spécifiques d'aider alors l'utilisateur à reformuler sa demande par l'intermédiaire d'une aide personnalisée (système de dialogue intelligent notamment). Le fait de passer par un formulaire SAV avec l'idée sous-jacente que les demandes seront traitées par un opérateur humain amènent probablement certains types d'utilisateurs à (re)formuler avec davantage d'informations pertinentes que lors de sa requête. Notamment d'autres termes sémantiquement proches sont utilisés pour formuler le même besoin informationnel, ils pourraient être alors utilisés pour des extensions de requête. A l'instar des dialogues homme-machine, des pistes d'améliorations des SRI seraient alors envisageables ; des dialogues évolutifs en fonction du profil d'utilisateur (détecté) permettraient de construire un cadre et une structure qui fourniraient des réponses plus appropriées au besoin informationnel de l'utilisateur.

BIBLIOGRAPHIE

[ADA 90] ADAM Jean-Michel. *Éléments de linguistique textuelle : théorie et pratique de l'analyse textuelle*. Liège (Belgique) : Mardaga, 1990, 234 p.

[AGR 95] AGRAWAL Rakesh, SRIKANT Ramakrishnan. *Mining Sequential Patterns*. In : Proceeding ICDE'95, 1995, pp.3-14.

[AIM 10] AIME Xavier, FURST Frédéric, KUNTZ Pascale, TRICHET Francky. *Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées*. In : EGC ?2010, [8 p.]

[ALL 02] ALLAN James et al. *Challenges in Information Retrieval and Language Modeling*. Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachussetts, Amherst, 2003.

[ALL 97] ALLEN B. *Information seeking in context*. In : P. Vakkari, R. Savolainen, B. Dervin (Eds.), Information seeking in context. Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts, 14-16 August 1996, Tampere, Finland (pp. 111-122). London, UK : Taylor Graham.

[AMI 02] AMIEL Virgine, TERRIER Patrice, POULAIN Gérard, CELLIER Jean-Marie. *Construction et évolution des représentations mentales lors des premières utilisations d'un service de dialogue en langage naturel*. In : Proceedings of the 14th French-speaking conference on Human-computer interaction (Conférence Francophone sur l'Interaction Homme-Machine), Poitiers, 2002. New York : ACM International Conference Proceeding Series, vol. 32, 2002, pp. 49-56.

[ANI 03] ANICK Peter. *Using terminological feedback for web search refinement : a log-based study*. In : Proceeding SIGIR'03, 26th annual international ACM SIGIR Conference on Research and Development in Informaion Retrieval, 2003, pp. 88-95

[ARG 09] ARGUELLO Jaime, MELLON Carnegie, DIAZ Fernando, CALLAN Jamie, CRESPO Jean-Francois. *Sources of evidence for vertical selection*. In : Proceeding SIGIR'09, 32nd international ACM SIGIR conference on Research and Development in Information Retrieval, 2009, pp. 315-322.

[ARM 93] ARMBRUSTER Bonnie B., ARMSTRONG James O. *Locating information in text : A focus on children in the elementary grades*, Contemporary Educational Psychology, 1993, vol. 18, pp. 139-161.

[BAI 08] BAI Jing, NIE Jian-Yun. *Adapting information retrieval to query contexts*, Information Processing and Management, vol. 44, issue 6, 2008, pp. 1901-1922.

[BAL 00] BALICCO Laurence, BEN ALI Salaheddine, PONTON Claude et Pouchot Stéphanie. *Apports de la génération automatique de textes en langue naturelle à la recherche d'informations*. In : ASCI'2000 : les dimensions d'une science de l'information globale, 28e congrès annuel de l'association canadienne des sciences de l'information, Edmonton, Alberta (Canada), 28-30 mai 2000, [s.l.] : 2000, [10 p.]

[BAO 07] BAO Shenghua, XUE Guirong, WU Xiaoyuan, YU Yong, FEI Ben, SU Zhong. *Optimizing web search using social annotations*. In : Proceeding WWW '07 Proceedings of the 16th international conference on World Wide Web, 2007, pp. 501-510.

[BAR 08] BARR Cory, JONES Rosie, REGELSON Moira. *The linguistic structure of English web-search queries*. In : Proceeding EMNLP'08, Conference on Empirical Methods in Natural Language Processing, 2008, pp. 1021-1030.

[BAR 94] BARKEMA Hank. *Determining the syntactic flexibility of idioms*. In : Creating and using English language, U. Fries, G. Tottie, P. Schneider (eds.), corpora (pp. 39-52). Amsterdam : Rodopi.

[BAU 08] BAUDOIN Frédéric. *Personnalisation des systèmes de dialogue en langage naturel*. Thèse de doctorat, Paris 6, 2008, 224 p.

[BEL 11] BELLOT Patrice (sous la dir. de). *Recherche d'information contextuelle, assistée et personnalisée*. Paris : Lavoisier, 2011, 302 p.

[BEL 08] BELKIN Nicholas J. *Some(what) Grand Challenges for Information Retrieval*. SIGIR Forum, vol. 42, n.1, 2008, pp.47-54.

[BEL 03] BELKIN Nicholas J., KELLY Diane Frances, KLIM Giyeong, KIM JaYoung, LEE HyukJin, MUSERAN Gheorghe, TANG Muh Chyun (Morris), YUAN Xiaojun, COOL Colleen. *Query Length in Interactive Information Retrieval*. In : Pro-

ceeding SIGIR'03, 26th annual international ACM SIGIR Conference on Research and Development in Informaion Retrieval, 2003, pp.205-212.

[BEL 95] BELKIN Nicholas J., COOL Colleen, STEIN Adelheit, THIEL Ulrich. *Cases, Scripts, and Information-Seeking Strategies : On the Design of Interactive Information Retrieval Systems*, Expert Systems With Applications, 1995, vol. 9, Issue 3, pp. 379-396.

[BEL 93] BELKIN Nicholas J. *Interaction with text : Information retrieval as information-seeking*, Information Retrieval, 1993, pp. 55-66.

[BEL 85] BELKIN Nicholas J., VICKERY Alina. *Interaction in information systems : An review of research from document retrieval to knowledge-based systems*. Cambridge : University Press, 1985, 250 p.

[BEL 82a] BELKIN Nicholas J., ODDY Robert N., BROOKS Helen M. *Ask for information retrieval. Part I. Background and theory*, juin 1982, vol.38, n.2, pp.61-71.

[BON 79] BON Laurence. *La formulation des besoins en information dans les entreprises : étude linguistique et proposition pour une méthode de classement*, Thèse de Doctorat, Paris, 1979.

[BOU 09] BOULESNANE Sabrina, BOUZIDI Laïd. Formulation des besoins informationnels dans une activité complexe et dynamique : l'audit et le conseil en Système d'Information et Nouvelles Technologies, CIBIMA'2009, vol.10, pp.72-84.

[BOU 09b] BOUMECHAAL Hasna, ALLIOUA Sofiane, BOUFAIDA Zizette. *Conversion des Requêtes en Langage Naturel vers nRQL*. In : CIIA, volume 547 of CEUR Workshop Proceedings, CEUR-WS.org, 2009, [11] p.

[BOU 09c] BOUIDGHAGHEN Ourdia, TAMINE-LECHANI Lynda, BOUGHANEM Mohand. *Vers la définition du contexte d'un utilisateur mobile de système de recherche d'information*. IN : Proceeding UbiMob '09 Proceedings of the 5th French-Speaking Conference on Mobility and Ubiquity Computing, 2009, pp. 25-31.

[BOU 11] BOUBÉE Nicole, TRICOT André. *L'activité informationnelle juvénile*. Paris : Lavoisier Hermès, 2011, 300 p. (Coll. Systèmes d'information et organisations documentaires).

[BRE 04] BRETIER Philippe, LE BIGOT Ludovic, PANAGET Franck, SADEK David. *De la représentation de l'interlocuteur vers un modèle utilisateur formel pour le dialogue personne-machine*. In : Proceeding IHM 2004. Proceedings of the 16th conference on Association Francophone d'Interaction Homme-Machine, 2004, pp. 21-28.

[BRO 02] BRODER Andrei. *A taxonomy of web search*. In : SIGIR FORUM, 2002, vol.36, n.2, pp.3-10.

[BRU 07] BRUSILOVSKY Peter, MILLAN Eva. *User Models for Adaptive Hypermedia and Adaptive Educational Systems*. In : BRUSILOVSKY Peter, KOBSA Alfred, NEJDL Wolfgang (Eds.) : The adaptive web. Berlin : Springer-Verlag, 2007, pp. 3-53.

[BRU 05] BRUCE Harry. *Personal, anticipated information need*, information research, vol.10, n.3, avril 2005, [15 p.].

[BRU 03] BRUANDET Marie-France, CHEVALLET Jean-Pierre. *Utilisation et construction de bases de connaissances pour la Recherche d'Informations*. In :GAUSSIER Eric,STEFANINI Marie-Hélène (sous la dir. de). Assistance intelligente à la recherche d'informations. Paris : Lavoisier : Hermes Science publ., 2003, pp.85-118. (Traité des sciences et techniques de l'information).

[BRU 01] BRUSILOVSKY Peter. *Adaptive Hypermedia*, User Modeling and User-Adapted Interaction, 2001, vol.11, pp.87-110.

[BRU 97] BRUZA Peter, DENNIS Simon. *Query Reformulation on the Internet : Empirical Data and the Hyper- index Search Engine*. In : Proceedings of RIA097, Computer- Assisted Information Searching on Internet, Montreal 1997, [12 p.].

[BUD 00] BUDZIK Jay, HAMMOND Kristian J. *User interactions with every day applications as context for just-in-time information access*. In : Proceedings of the 5th international conference on Intelligent user interfaces, pp.44-51, 2000.

[BYS05] BYSTRÖM Katriina, HANSEN Preben. *Conceptual Framework for Tasks in Information Studies*. In : Journal of the American Society for Information Science and Technology, vol.56, issue 10, 2005, pp. 1050-1061.

[CAB 11] CABANAC Guillaume, CHEVALIER Max, CIACCIA A., CLAVEL Céline, HUBERT Gilles, JULIEN Christine, SOULÉ-DUPUY Chantal, TRICOT André. Recherche d'information et modélisation usagers. In P. Bellot (Ed.) Recherche d'information contextuelle, assistée et personnalisée. Paris : Hermès, 2011.

[CAE 03] CAELEN Jean. *Dialogue homme-machine et recherche d'information*. In : GAUSSIER Eric, STEFANINI Marie-Hélène (sous la dir. de). Assistance intelligente à la recherche d'informations. Paris : Lavoisier : Hermes Science publ., 2003, pp.219-254. (Traité des sciences et techniques de l'information).

[CAE01a] CAELEN Jean, ROUILLARD José. *Le système HALPIN : recherche documentaire en langue naturelle et dialogue multimodal*, Interaction Homme-Machine, 2001, vol.2, n.2, pp.55-72.

[CAL 06] CALABRETTO Sylvie, EGYED-ZSIGMOND Előd. *Recherche d'Information en Contexte*, École d'Automne RIA (EARIA'2006), 88 p.

[CAR 12] CARPINETO Claudio, ROMANO Giovanni. *A Survey of Automatic Query Expansion in Information Retrieval*. In : Journal ACM Computing Surveys (CSUR, 2012, vol. 44, n. 1, [56 p.].

[CAR 08] CARMAN Mark James, BAILLIE Mark, CRESTANI Fabio. *Tag data and personalized information retrieval*. In : Proceedings of the 2008 ACM Workshop on Search in Social Media, 26 October 2008 to 30 October 2008, Association for Computing Machinery. New York, 2008, pp. 27-34.

[CAS 12] CASE Donald O. *Looking for information : A survey of research on information seeking, needs and behavior*. San Diego (Etats-Unis) : Academic Press, 2002, 350 p.

[CEL 10] CELLIER Peggy ,CHARNOIS Thierry. *Fouille de données séquentielles d'item-sets pour l'apprentissage de patrons linguistiques*. In : Actes TALN'2010, Montréal, 19 ?23 juillet 2010, [6 p.].

[CHA 10] CHAKER Hamdi, CHEVALIER Max, SOULÉ-DUPUY Chantal, TRICOT André. Système de recherche d'information pour les tâches métier. CORIA'2010, Conférence en Recherche d'Information et Applications, 18-20 mars 2010, Sousse, Tunisie.

[CHA 96] CHANET Catherine. *La demande dans le dialogue finalisé : de la surface linguistique aux représentations de l'action*, thèse de doctorat, Grenoble 3, 1996, 450 p.

[CHA 02] CHAUDIRON Stéphane, IHADJADENE Madjid. *Quelle place pour l'usager dans l'Ševaluation des systèmes de recherche d'Šinformation ? Du paradigme système au paradigme cognitif* . In : Journée d'ŠInformation URBAMET, 5 mars 2002, (12 p.].

[CHA 04a] CHAUDIRON Stéphane. *L'évaluation des systèmes de recherche d'informations*. In :IHADJADENE Madjid (sous la dir. de). Les systèmes de recherche d'informations : modèles conceptuels. Paris : Hermès science publ. : Lavoisier, 2004, pp.(Traité des sciences et techniques de l'information)

[CHA 04b] CHAUDIRON Stéphane. *La place de l'usager dans l'évaluation des systèmes de recherche d'informations*. In : CHAUDIRON Stéphane (sous la dir. de). Evaluation des systèmes de traitement de l'information. Paris : Hermès science publ. : Lavoisier, 2004, pp.287-310

[CHA 05] CHAU Michael, FANG Xiao, LIU SHENG Olivia R. *Analysis of the Query Logs of a Web Site Search Engine*, Information of the American Society dor Information Science and Technology, vol. 56, n. 13, 2005, pp.1363-1376.

[CHE 12] LI Chenliang, WENG Jianshu, HE Qi, YAO Yuxia, DATTA Anwitaman, SUN Aixin, LEE Bu-Sung. *TwiNER : Named Entity Recognition in Targeted Twitter Stream*. In : Proceeding SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 721-730.

[CHU 03] CHUANG Shui-Lung, CHIEN Lee-Feng. *Automatic query taxonomy generation for information retrieval applications*, Online Information Review, 2003, vol. 27, issue 4, pp.243-255.

[CLE 67] CLEVERDON Cyril W. *The Cranfield tests on index language devices*, Aslib proceedings, 1967, vol.19, nř6, pp.173-193.

[COO 02] COOL Colleen,SPINK Amanda. *Issues of context in information retrieval (IR) : an introduction to the special issue*, Information Procesing Management, 2002, vol. 38, Issue 5, pp.605-611.

[COO 01] COOL Colleen. *The concept of situation in information science*. Annual Review of Information Science and Technology, 2001, vol. 35, pp. 5-42.

[COU 07] COURTRIGHT Christina. *Context in information behavior research*. Annual Review of Information Science and Technology, 2007, vol. 41, pp. 273-306.

[CRO 02] CRONEN-TOWNSEND Steve, ZHOU Yun, CROFT W. Bruce. *Predicting Query Performance*, Proceedings of the 25th annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, 2002, pp.299-306.

[DAO 09] DAOUD Mariam. *Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche*. Thèse de doctorat soutenue publiquement le 10 Décembre 2009, Université Paul Sabatier - Toulouse III, 2009, 287 p.

[DAO 09b] DAOUD Mariam, TAMINE-LECHANI Lynda, BOUGHANEM Mohan. A SESSION BASED PERSONALIZED SEARCH USING AN ONTOLOGICAL USER PRO ?LE. In : Proceeding SAC '09 . Proceedings of the 2009 ACM symposium on Applied Computing, 2009, pp. 1732-1736.

[DEGROC 13] DE GROC Clément. *Collecte orientée sur le Web pour la recherche d'information spécialisée.*, Thèse de Doctorat, Université Paris Sud, Juin 2013.

[DEL 05] DELECROIX Bertrand, EPPSTEIN Renaud. *Une analyse des requêtes d'un moteur intranet - Vers une amélioration du système d'information*. In : Journée sur les systèmes d'information élaborée, île Rousse 2005, 2005, [8 p.].

[DEL 00] DE LOUPY Claude, BELLOT Patrice. *Evaluation of Document Retrieval Systems and Query Difficulty*, Actes du LREC 2000 Satelitte Worshop, Using Evaluation within HLT Programs, 2000, pp.34-40.

[DON 07] DONG Guozhu, PEI Jian. *Sequence Data Mining*. New-York : Springer, 2007, 168 p.

[DUB 01] DUBOIS Jean. *Dictionnaire de linguistique et des sciences du langage*. Paris : Larousse, 2001, 514 p. (Trésors du français)

[ELL 97] ELLIS David, HAUGAN Merete. *Modelling the information seeking patterns of engineers and research scientists in an industrial environment*, Journal of Documentation, 1997, vol.53, n.4, pp.384-403.

[ELL 92] ELLIS David. *Paradigms and proto-paradigmes in information retrieval research*. In : VAKKARI Pertti, CRONIN Blaise (sous la dir. de). Conceptions of library and information science : historical, empirical, and theoretical perspectives. Londres (Royaume-Uni) : Taylor Graham, 1992, pp 165-186.

[FAY 97] FAYOL Michel. *Des idées au texte : psychologie cognitive de la production verbale, orale et écrite*. Paris : Presses Universitaires de France (PUF), 1997, 296 p.

[GAR 10] GARCIA-FERNANSDEZ Maria. Génération de réponses en langue naturelle orales et écrites pour les systèmes de question-réponse en domaine ouvert.Thèse de Doctorat, Université Paris Sud 11 Orsay, 2010, 201 p.

[GAU 97] GAUSSIER Eric, GREFENSTETTE Grégory, SCHULZE Maximilian Bruno. *Traitement du langage naturel et recherche d'informations : quelques expérience sur le français.* In : Actes des premières journées scientifiques et techniques FRANCIL'97. Avignon : AUPELF-UREF, 1997, pp.9-14.

[GOK 08] GÖKER Ayse, MYRHAUG Hans. *Evaluation of a mobile information system in context*, Journal Information Processing and Management : an International Journal, vol. 44, Issue 1, 2008, pp.39-65.

[GOK 02] GÖKER Ayse, MYRHAUG Hans. *User Context and Personalisation*. In : 6th European Conference on Case Based Reasoning, 4-7 September 2002, Aberdeen, Scotland, [7 p.].

[GOL 11] GOLLAPUDI Sreenivas, IEONG Samuel, NTOULAS Alewandros, PAPARIZOS Stelios. *Efficient query rewrite for structured web queries*. In : Proceeding CIKM '11, 20th ACM international conference on Information and knowledge management, pp. 2417-2420.

[GUO 08] GUO Jiafeng, XU Gu, CHENG Xueqi. *A Unified and Discriminative Model for Query Refinement*. In : Proceeding SIGIR'08, SIGIR Conference on Research and Development in Information Retrieval, 2008, pp.379-386.

[GUO 09] GUO Jiafeng, XU Gu, CHENG Xueqi, LI Hang. *Named entity recognition in query*. In : Proceeding SIGIR'09, 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp 267-274.

[GUT 88] GUTHRIE John T. *Locating Information in Documents : Examination of a Cognitive Model* , Reading Research Quarterly, 1988, vol. 23, n. 2, pp. 178-199. Mahwah (Etats-Unis) : Lawrence Erlbaum Associates, 1999, pp. 195-218.

[HAB 98] HABERT Benoit. *Des mots complexes possibles aux mots complexes existants : l'apport des corpus*, Habilitation à Diriger des Recherches (HDR), Université Lille 3, 1998, 185 p.

[HAC 98] HACKOS Joann T, REDISH Janice C. *User and Task Analysis for Interface Design*. New York : John Wiley and Sons, Inc., 1998, 512 p.

[HAR 06] HARRATHI Rami, CALABRETTO Sylvie . *Un modèle de qualité de l'information*. In : G. Ritschard and C. Djeraba, editors, EGC, volume RNTI-E-6 of Revue des Nouvelles Technologies de l'Information, pp. 299-304. Cépadu ditions, 2006.

[HAR 52] ZELLIG Harris. *Discourse Analysis, Language*, 1952, n.28, vol.1, pp.1-30.

[HAR 69] ZELLIG Harris. *The two systems of Grammar : Report and Paraphrase*. In : Papers in structural and transformational linguistics. Dordrecht : D. Reidel.

[HAR 96] HARTER Stephen. *Variations in relevance assessments and the measurement of retrieval effectiveness*, JASIS, 1996, vol. 47, n. 1, pp. 37-49.

[HE 08] HE Jiyin, LARSON Martha, DE RIJKE Maarten. *Using Coherence-Based Measures to Predict Query Difficulty*. In : Proceeding ECIR'08 Proceedings of the IR research, 30th European conference on Advances in information retrieval, 2008, pp.689-694

[HE 02] HE Daging, GOKER Ayse, HAPER David J. *Combining evidence for automatic web session identification*, Information Processing and Management, 2002, vol. 38, n. 5, pp. 727-742.

[HEI 08] HEINECKE Johannes, SMITS Grégory, CHARDENON Christine, GUIMIER DE NEEF Emilie, MAILLEBUAU Estelle, BOUALEM Malek. *TiLT : plate-forme pour le traitement automatique des langues naturelles*, revue TAL, 2008, vol.49, n.2.

[HEY 08] HEYMANN Paul, KOUTRIKA Georgia, GARCIA-MOLINA Hector. *Can social bookmarking improve web search ?*. In : Proceeding WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008, pp. 195-206.

[HI 02] HE Daqing, GOKER Ayse, HARPER David J. *Combining Evidence for Automatic Web Session on Web Search Engine*, Journal Information Processing and Management, 2002, vol. 38, n. 5,pp.727-742.

[HON 01] HONE Kate, BABER Chris. *Designing habitable dialogues for speech-based interaction with computers*, International Journal of Human-Computer Studies, 2001, vol. 54, issue 4, pp. 637-662.

[HUA 10] HUANG Jian. *Exploring web scale language models for search query processing*. In : Proceeding WWW '10, 19th international conference on World wide web, 2010, pp. 451-460.

[HUA 09] HUANG Jeff, EFTHIMIADIS Efthimis N. *Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs*. In : Proceeding CIKM'09, 18th ACM conference on Information and knowledge management, 2009, pp.77-86.

[HUP 06] HUPFER Maureen E., DETLOR Brian. *Gender and Web Information Seeking : A Self-Concept Orientation Model*, Journal of the American Society for Information Science and Technology, 2006, pp. 1105-1115.

[IHA 99] IHADJADENE Madjid. *La recherche et la navigation dans un système de recherche dŠinformation grand public : Le cas des hypercatalogues sur lŠInternet*, thèse de doctorat, Lyon 1, 1999, 281 p.

[ING 05] INGWERSEN Peter, JARVELIN Kalervo. *The Turn. Integration of information seeking and retrieval in context*. Dordrecht : Springer, 2005, 448 p.

[ING 96] INGWERSEN Peter. *Cognitive perspectives of information retrieval interaction : elements of a cognitive IR theory*, Journal of Documentation, 1996, vol. 52, nř 1, pp. 3-50.

[ING 92] INGWERSEN Peter. *Information retrieval interaction*. London : Taylor Graham, 1992, 246 p.

[JAC 00] JACQUEMIN Christian, ZWEIGENBAUL Pierre. *Traitement automatique des langues pour l'accès au contenu des documents*. In : LE MAITRE Jacques, CHARLET Jean, GARBAY Catherine. Traitement automatique des langues pour l'accès au contenu des documents. Cepadues : Toulouse, 2000, pp.71-109.

[JAC 97] JACQUEMIN Christian. *Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus.* Habilitation à Diriger des Recherches (HDR), Université de Nantes, 1997.

[JAN 08] JANSEN Bernard J.,BOOTH L. Danielle, SPINK Amanda. *Determining the informational, navigational and transactional intent of Web queries*, Information Processing and Management, 2008, vol. 44, pp. 1251-1266.

[JAN 07] JANSEN Bernard J.,SPINK Amanda, BLAKELY Chris, KOSHMAN Sherry. *Defining a session on Web search engines*, Journal of the American Society for Information Science and Technology, 2007, vol. 58, n.6, pp. 862-871.

[JAN 06] JANSEN, Bernard J. *Search Log Analysis : What It Is, What's Been Done, How to Do It*. Library and Information Science Research, 2006, vol. 28, n. 3, pp. 407-432.

[JAN 05] JANSEN, Bernard J., SPINK Amanda, PEDERSEN Jan. *A temporal Comparaison of AltaVista Web Searching*, Journal of the American Society for Information Science and Technology, 2005, vol. 56, n. 6, pp.559-570.

[JAN 00a] JANSEN Bernard J., SPINK Amanda, SARACEVIC Tefko. *Real life, real users, and real needs : A study ans analysis of user queries on the Web*, Information Processing and Management, 2000, n. 36, vol.2, pp.207-227.

[JON 07] JONES William. HOW PEOPLE KEEP AND ORGANIZE PERSONAL INFORMATION. In : Personal Information Management. Jones W. and Teevan J. (Eds.). Seattle, Washington : University of Washington Press, 2007, pp.35-56.

[JON 05] JONES William,JIRANIDA PHUWANARTNURAK Ammy,GILL Rajdeep ,BRUCE Harry. Don ?t Take My Folders Away ! Organizing Personal Information to Get Things Done. In : Proceeding CHI EA '05 CHI '05 Extended Abstracts on Human Factors in Computing Systems. New York : ACM : 2005, pp. 1505-1508.

[JUL 04] JULIEN Heidi E., MICHELS David. *Intra-individual information behavior in daily life*, Information Processing and Management, 2004, vol. 40, Issue 3, pp. 547-562.

[KAM 07] KAMVAR Maryam, BALUJA Shumeet. *Deciphering Trends In Mobile Search*, Journal Computer, 2007, vol. 40, Issue 8, pp.58-62.

[KAN 05] KANG In-Ho. *Transactional query identification in web search*. In AIRS'05 : Proceedings Information Retrieval Technology, Second Asia a Information Retrieval Symposium, Jeju Island, Korea, 2005, pp. 221-232.

[KAN 03] KANG In-Ho, KIM GilChang. *Query Type Classification for Web Document Retrieval*. In : Proceeding SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. New York : ACM, 2003, pp.64-71.

[KHO 09] KHOUSSAINOVA Nodira, BALAZINSKA Magdalena, GATTERBAUER Wolfgang, KWON YongChul, SUCIU Dan. *A case for a collaborative query management system.* In : 4 th Biennial Conference on Innovative Data Systems Research (CIDR) January 4-7, 2009, Asilomar, California, USA

[KIM 12] KIM Jin Young, CROFT Bruce C. *A field relevance model for structured document retrieval*. In : Proceeding ECIR'12, 34th European conference on Advances in Information Retrieval, 2012, pp. 97-108.

[KIM 09] KIM Jin Young, XUE Xiaobing, CROFT Bruce C. *A Probabilistic Retrieval Model for Semistructured Data*. In : Proceeding ECIR '09, 31th European Conference on IR Research on Advances in Information Retrieval, 2009, pp. 228-239.

[KIM 08] KIM Kyung-Sun. *Effects of emotion control and task on Web searching behavior*, Information Processing and Management, vol. 44, Issue 1, 2008, pp.373 ?385.

[KIR 08] KIRCHHOFF Lars, STANOEVSKA-SLABEVA Katarina, NICOLAI Thomas, FLECK Matthes. USING SOCIAL NETWORK ANALYSIS TO ENHANCE INFORMATION RETRIEVAL SYSTEMS. In : 5th Conference on Applications of Social Network Analysis (ASNA), Zurich, 2008. September 2008, [21 p.].

[KRO 92] KROVETZ Robert, CROFT Bruce W. *Lexical Ambiguity and Information Retrieval*, ACM Transactions on Information Systems, 1992, vol 10, n.2, pp.115-141.

[KUH 91] KUHLTHAU Carol Collier. *Inside the search process : information-seeking from the user's perpective*, Journal of the American Society for Information Science (JASIST), 1991, vol. 42, n.5, pp. 361-371.

[KUM 08] KUMARAN Giridhar, ALLAN James. *Adapting information retrieval systems to user queries, Information Processing and Management*, 44, vol.6, pp.1838-1862.

[LAI 99] LAINE-CRUZEL Sylvie. *Profildoc : filtrer une information exploitable*, BBF, 1999, t. 44, n.5, pp.60-64.

[LAL 11] LALLEMAN Fanny. *Analyse de l'ambiguïté des requêtes utilisateurs par catégorisation thématique*. In : RECITAL'2011, [10 p.]

[LAN 08] LANG Hao Lang, WANG Bin, JONES Gareth, LI Jin-Tao, DING Fan, LIU Yi-Xua. *Query performance prediction for information retrieval based on covering topic score*, Journal of Computer Science and Technology, 2008, vol 23, n.4, pp.590-601.

[LAT 13a] LATOUR Marilyne. *Expressions différenciées des besoins informationnels en Langue Naturelle : une contextualisation par la tâche*. In : Actes CEC-TAL'2013, Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage Naturel et ses applications, 26-27 septembre 2013, Montréal (Canada), [8 p.], à paraître.

[LAT 13b] LATOUR Marilyne, DANESI Charlotte. *Extraction automatique de données économiques : un exemple d'application chez ReportLinker*. In : Actes RJCIA'2013, Rencontres des Jeunes Chercheurs en Intelligence Artificielle, 1-5 Juillet 2013, Lille (France), [8 p.], à paraître.

[LAT 05] LATOUR Marilyne. Recherche d'informations techniques : du besoin d'informations à la formulation des requêtes, Mémoire de master 2 en Sciences de l'information et de la communication, option recherche : informatique, discours et documents, sous la dir. d'Evelyne MOUNIER et de Céline PAGANELLI. Grenoble : Université Stendhal Grenoble 3, 2005, 83 p.

[LAU 99] LAU Tessa, HORVITZ Eric. *Patterns of search : analyzing and modeling web query refinement*. In UM'99 : Proceedings of the seventh international conference on User modeling. Springer-Verlag, 1999, pp. 119-128.

[LEC 98] LE COADIC Yves-François. *Le besoin d'information : formulation, négociation, diagnostic*. Paris : ADBS Editions, 1998, 191 p.

[LEC 96] LECKIE Gloria.J., PETTIGREW Karen E., SYLVAIN Christian. *Modelling the information seeking of professionals : a general model derived from research on engineers, health care professionals and lawyers*, Library Quarterly, 1996, vol. 66, Issue 2, pp. 161-193.

[LEG 06] LE BIGOT Ludovic, JAMET Eric,ROUET Jean-François, POULAIN Gérard. *Ordre des informations et effet de modalité pour une recherche de restaurants*. In ACM International Conference Proceeding Series : Proceedings of the 18th international conference on Association Francophone d'Interaction Homme-Machine IHM '06. Montréal, Québec (Canada) : ACM Press, 2006, vol. 133, pp. 83-89.

[LEG 05] LE BIGOT Ludovic, JAMET Eric, ROUET Jean-François. Ordre des informations dans la formulation d'une recherche d'information. In : Raufaste Eric, Tricot André (Eds.), Troisième Journées d'étude en Psychologie ergonomique EPIQUE'05. Toulouse : Université de Toulouse, 2005, pp. 149-155 .

[LEI 96] LEVENSHTEIN Vladimir I. *Binary codes capable of correcting deletions, insertions, and reversals*, Soviet Physics Doklady, 1996, vol. 10, n.8., pp. 707-710.

[LEV 13a] LEVA Simon. *Les sessions de recherche comme contexte des requêtes*. In : Actes EGC'2013, Toulouse, 2013, [12] p.

[LEV 13b] LEVA Simon, FAESSEL Nicolas. *Détection automatique des sessions de recherche par similarité des résultats provenant d'une collection de documents externe*. In : Actes TALN-RECITAL'2013, Les Sables d'Olonne, pp.217-230.

[LI 10] LI Xiao. *Understanding the Semantic Structure of Noun Phrase Queries*. In : Proceeding ACL'10, 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 1337-1345

[LI 09] LI Xiao, WANG Ye-Yi, ACERO Alex. *Extracting Structured Information from User Queries with Semi-Supervised Conditional Random Fields*. In : Proceeding SIGIR'09, 2009, pp. 572-579.

[LI 08] LIA Yuelin, BELKIN Nicholas J.*A faceted approach to conceptualizing tasks in information seeking*, Information Processing and Management, 2008, vol. 44, n. 6, pp. 1822-1837.

[LI 08b] LI Xiao, WANG Ye-Yi,ACERO Alex. *Learning Query Intent from Regularized Click Graph*. In :Proceedings of SIGIR ?08, 31st Conference on Research and Development in Information Retrieval, 2008, pp.339-346.

[LI 03] LI Xiaoyan, CROFT Bruce W. *Time-based language models*. In CIKM'03 : Proceedings of the twelfth international conference on Information and knowledge management, 2003, pp. 469 ?475.

[LI 02] LI Xin, ROTH Dan. *Learning question classifiers*. In : Proceeding COLING'02, 19th international conference on Computational linguistics, 2002, vol.1 , pp.1-7.

[LIT 02] LITMAN Diane J., PAN Shimei. *Designing and Evaluating an Adaptive Spoken Dialogue System*, UMUAI (User Modeling and User-Adapted Interaction), 2002, vol.12, pp.111-137.

[LOU 04] DE LOUPY Claude, CRESTAN Eric. *Système de recherche d'informations et traitement du langage naturel*. In : IHADJADENE Madjid (sous la dir. de). Les systèmes de recherche d'informations : modèles conceptuels. Paris : Hermès science publ. : Lavoisier, 2004, pp. 139-158. (Traité des sciences et techniques de l'information)

[MCC 99] MCCREADIE Maureen, RICE Ronald E. *Trends in analyzing access to information*. Part I : Cross-disciplinary conceptualizations of access. Information Processing and Management, 1999, vol. 35, Issue 1, pp. 45-76.

[MAN 09] MANSHADI Medhi, LI Xiao. *Semantic Tagging of Web Search Queries*. In : Proceeding ACL'09 , 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, vol.2, pp.861-869.

[MAN 02] MANDL Thomas, WOMSER-HACKER Christa. *Linguistic and Statistical Analysis of the CLE Topics*, CLE Workshop, 2002, [8 p.].

[MAT 01] MATLIN Margaret W. *la cognition : une introduction à la psychologie cognitive*. Paris : De Boeck Université, 2001, 749 p.

[MAR 76] MARCELLESI Christiane. *L'expression linguistique du référent : la dénomination* ; Langue française, 1976, n. 32, pp. 40-62.

[MEN 83] MENDELSOHN Patrick. *Réflexions sur l'objet d'étude de la pensée naturelle*. In : La Pensée Naturelle : structures, procédures et logique du sujet. Paris : PUF, 1983, pp. 295-311.

[MEN 09] MENDOZA Marcelo, ZAMORA Juan. *Building Decision Trees to Identify the Intent of a User Query*. In : Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KESŠ09), Santiago, Chile, pp. 285-Ű292.

[MEY 11] MEYER Franck,GAUSSIER Eric, CLEROT Fabrice, SCHLUTH Julien. *Apport des données thématiques dans les systèmes de recommandation hybridation et démarrage à froid*. In : EGC'2011, pp.215-220.

[MIL 90] MILLER George A., BECKWITH Richard, FELLBAUM Christiane, GROSS Derek, MILLER Katherine J. *Introduction to WordNet : An online lexical database*, International Journal of Lexicology, 1990, vol.3, n.4, pp.235-244.
DE MONTMOLLIN Maurice D. *LŠergonomie*. Paris : La Découverte, 2007, 120 p. (Repères)

[MOR 02] MORIZIO Claude (sous la dir. de). *La recherche d'information*. Paris : ADBS : A. Colin, 2004, 126 p. (128. Information, documentation)

[MOT 13] MOTHE Josiane. *Recherche d'Information et Traitement Automatique des Langues Naturelles* . In : Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013), 2013, Les Sables d'Olonne, France, [p.2].

[MOT 11] MOTHE Josiane. *Recherche d'information contextuelle : le cas des requêtes*. In : Recherche d'information et modélisation usagers. In P. Bellot (Ed.) Recherche d'information contextuelle, assistée et personnalisée. Paris : Hermès, 2011.

[MOT 05] MOTHE Josiane, TANGUY Ludovic. *Linguistics features to predict query difficulty - A case study on previous TRES campaign*. In : ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications

[NAD 07] NADEAU David, SEKINE Satoshi. *A survey of named entity recognition and classification*, Lingvisticae Investigationes, 2007, vol. 30, n. 1, pp. 3-26.

[NAI 98] NAIT-BAHA Leïla, JACKIEWICZ Agata, LAUBLET Philippe. *Reformulation des requêtes et extraction de phrases pertinentes pour la collecte d'informations sur le Web*. In : Actes RIFRA'98 (Rencontre Internationale sur l'Extraction le Filtrage et le Résume Automatique), 1998, pp 177-190.

[NAM 00] NAMER Fiammetta. FLEM *: un analyseur flexionnel du français à base de règles*, TAL, 2000, 41-2, pp.523-547.

[NAU 98] NAULLEAU Elie. *Apprentissage et filtrage syntaxico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire*, Thèse de Doctorat en Informatique, Université Paris XIII Villetaneuse, 1998, 196 p.

[NAV 99] NAVARRO-PRIETO Raquel, SCAIFE Mike, ROGERS Yvonne. *Cognitive strategies in web searching*. In : proceedings of the 5 th Conference on Human Factors and the Web, Gaithersburg (Maryland, Etats-Unis) 5 juillet 1999. NIST (National Institute of Standards and Technology), 1999, [12 p.]

[NIA 13] NIANG Cheikh, BOUCHOU Béatrice, LO Moussa, SAM Yacine. *Ré-écriture de requêtes dans un système d'intégration sémantique*. In : Proceedings EGC 2013 Toulouse, 2013, pp.383-388

[NIE 03] NIE Jian-Yun. *Le domaine de recherche d'information : un survol d'une longue histoire*. In : GAUSSIER Eric, STEFANINI Marie-Hélène (sous la dir. de). Assistance intelligente à la recherche d'informations. Paris : Lavoisier : Hermes Science publ., 2003, pp.19-28. (Traité des sciences et techniques de l'information)

[PAP 09] PAPARIZOS Stelios, NTOULAS Alexandros, SHAFER John, AGRAWAL Rakesh. *Answering web queries using structured data sources*. In : Proceeding SIGMOD '09, ACM SIGMOD International Conference on Management of Data, 2009, pp.1127-1130.

[PAS 08] PASCA Marius, VAN DURME Benjamin. *Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs*. In : Proceedings of ACL'08, 2008, pp. 19-27.

[PAS 07] PASCA Marius, VAN DURME Benjamin. *What you seek is what you get : extraction of class attributes from query logs*. In : Proceeding IJCAI'07, 20th international joint conference on Artifical Intelligence, 2007, pp. 2832-2837.

[PAS 06] PASS Greg, CHOWDHURY Abdur, TORGESON Cayley. *A picture of search*. In : Proceeding InfoScale'06, 1st international conference on Scalable Information Systems, 2006, Article n.1 , [7 p.]

[PAT 02] PATERNOSTRE Marjorie, FRANCQ Pascal , LAMORAL Julien, WARTEL David, SAERENS Marco. *Carry, un algorithme de désuffixation pour le Français*, Projet GALILEI (Generic Analyser and Listener for Indexed and Linguistics Entities of Information), Juillet 2012, [17 p.]

[PHA 07] PHAN Nina, BAILEY Peter, WILKINSON Ross. *Understanding the Relationship of Information Need Specificity to Search Query Length*. In : Proceeding SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. New York : ACM, 2007, pp. 709-710.

[PIE 91] PIERREL Jean-Marie, SABAH Gérard. *Dialogue homme-machine en langage naturel ecrit et oral : bilan des approches du CRIN et du LIMSI*. In : Actes des 2èmes journées nationales du GRECO PRC Communication homme-machine, EC2 editeur, Toulouse, janvier 1991, pp. 91-111.

[PLU 11] PLU Michel, HEINECKE Johannes. *Interprétation linguistique de requêtes pour un moteur de questions réponses grand public*. In : Actes CORIA'2011, [8 p.]

[POI 11] POIBEAU Thierry. *Traitement automatique du contenu textuel*. Paris : Hermès Lavoisier, 2011, 222 p.

[POL 00] POLITY Yolla. *L'évolution des paradigmes dans le domaine de la recherche d'information*. Communication au groupe de travail "Théories et Pratiques Scientifiques" (TPS) de la Société Française des Sciences de l'Information et de la Communication (SFSIC), 3 mars 2000.

[POR 80] PORTER Martin. *An algorithm for suffix stripping*. In : Program : electronic library and information systems, 1980, vol. 14, Iss : 3, pp.130-137.

[PU 02] PU Hsiao-Tieh, CHUANG Shui-Lung, YANG Chyan. *Subject categorization of query terms for exploring Web users' search interests*, JASIST, Journal of the American Society for Information Science and Technology, 2002, vol. 53, issue 8, pp. 617-630.

[QUI 13] QUINTANA Manon. *Inbenta Semantic Search Engine : un moteur de recherche sémantique inspiré de la Théorie Sens-Texte*. In : In : Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013), 2013, Les Sables d'Olonne, France, pp. 791-792.

[RAG 95] RAGHAVAN Vijay V.SEVER Hayri. *On the reuse of past optimal queries*. In : Proceeding SIGIR '95 Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, 1995, pp. 344-350.

[RAM 06] RAMIREZ Goergina , DE VRIES Arjen P.. *Relevant contextual features in XML retrieval*. In : Proceedings of the 1st international conference on Information Interaction in Context. New York : ACM, 2006, pp. 95-110.

[RAY 13] RAYNAL Céline. *L'apport du TAL dans la catégorisation d'évènements de sécurité aéronautiques*. In : Actes du colloque Linguistique et Traitement Automatique des Langues pour l'Aéronautique et l'Espace. Université Toulouse II- Le Mirail, 2 Juillet 2013, p.4.

[RIJ 79] VAN RIJSBERGEN Cornelis Joost. *Information Retrieval* , Butterworths, 1979.

[ROD 06] RODE Henning, HIEMSTRA Djoerd.*Using Query Profiles for Clarification*. In : Advances in Information Retrieval. 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006. Proceedings, 2006, pp 205-216.

[ROS 04] ROSE Daniel, LEVINSON Danny. *Understanding User Goals in Web Search*. In : Proceeding WWW '04 Proceedings of the 13th international conference on World Wide Web. New York : ACM, 2004, pp.13-19.

[ROS 94] ROSSARI Corinne. *Les opérations de reformulation : analyse du processus et des marques dans une perspective contrastive français-italien*. Berne : Peter Lang, 1994.

[ROU 01] ROUET Jean-François. *Les activités documentaires complexes : aspects cognitifs et développementaux,* Mémoire pour l'Habilitation à Diriger des Recherches (HDR), Université de Poitiers, 2001,184 p.

[ROU 98] ROUET Jean-François, TRICOT André.*Chercher de l'information dans un hypertexte : vers un modèle des processus cognitifs*, Les hypermédias : approches cognitives et ergonomiques 1998, HS, pp.57-74.

[SAN 10] SANTAMARIA Celina, GONZALO Julio, ARTILES Javier *Wikipedia as sense inventory to improve diversity in web search results*. In : ACL, 2010, pp.1357-1366.

[SAN 08] SANDERSON Mark. *Ambigous queries : test collections need more sense*. In : SIGIR'08, pp.499-506.

[SAN 00] SANDERSON Mark. *Retrieving with good sense*, Information Retrieval, 2000, vol.2, n.1, pp.45-65.

[SAN 94] SANDERSON Mark. *Word sense disambiguiation and information retrieval*. In : Actes 17th annual international ACM-SIGIR. Conference on research and development in information retrieval. New York : ACM Press, 1994, pp.142-151.

[SAR 10] SARKAS Nikos, PAPARIZOS Stelios. *Structured annotations of web queries*. In : Proceeding SIGMOD '10, ACM SIGMOD International Conference on Management of data, 2010, pp. 771-782.

[SAR 97] SARACEVIC Tefko. *The stratified model of information retrieval interaction : Extension and applications*, Proceedings of the American Society for Information Science, nř34, pp.313-327.

[SAR 97b] SARACEVIC Tefko. *Users lost : reflections of the past, future and limits of information science*. In : SIGIR Forum, 1997, vol. 31, n.2, pp. 16-27.

[SAR 96] SARACEVIC Tefko. *Modeling interaction in Information Retrieval (IR) : a review and proposal*, Proceedings of the American Society for Information Science, 1996, vol.33, pp.3-9.

[SAV 12] SAVOLAINEN Reijo. *Conceptualizing information need in context*, Information Research, 2012, vol. 17, no. 4, [13 p.].

[SCH 94] SCHAMBER Linda. *Relevance and information behaviour*, Annual Review of Information Science and Technology (ARIST), vol. 29, pp.3-48.

[SCH 95] SCHMID Helmut. *Improvements in Part-of-Speech Tagging with an Application to German*. In : Proceedings of the ACL SIGDAT-Workshop, 1995. Dublin, Ireland.

[SER 05] SERRES Alexandre. *Les sept grandes tendances de la recherche d'Šinformation sur Internet*, lettres de l'ŠUrfist, n.34, 2005, pp.15-18.

[SHE 06] SHEN Dou ,SUN Jian-Tao ,YANG Qiang ,CHEN Zheng.*Building bridges for web query classification*. In : Proceeding SIGIR '06, 29th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, pp.131-138.

[SHE 05] SHEN] XUEHUA, TAN BIN, ZHAI CHENGXIANG. *Context-Sensitive Information Retrieval Using Implicit Feedback*, INTERNATIONAAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL,2005, pp.43-50.

[SHE 03] SHEN] XUEHUA, ZHAI CHENGXIANG. *Exploiting query history for document ranking in interactive information retrieval*, INTERNATIONAAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 2003, pp. 277-278.

[SHI 92] SHIFRONI Eyal, SHANON Benny. *Interactive user modeling : An integrative explicit-implicit approach*, UMAI (User Modeling and User-Adapted Interaction), 1992, vol. 2, issue 4, pp. 331-365.

[SHN 05] SHNEIDERMAN Ben, PLAISANT Catherine. *Designing the User Interface : Strategies for Effective Human-Computer Interaction*, 4e ed. Boston (Etats-Unis) : Addison Wesley, , 2004, 672 p.

[SIL 99] SILVERSTEIN Craig, HENZINGER Monika, MARAIS Hannes, MORICZ Michael. *Analysis of a very large web search engine query log*, SIGIR Forum, vol.33 n.1, pp. 6-12, 1999.

[SIM 06] SIMONNOT Brigitte. *Le besoin d'information : principes et compétences*. In : Actes de la journée thémat'IC : Information, besoins et usages, 17 mars 2006, Strasbourg-Illkirch, présentation powerpoint, [9 p.]

[SOH 08] SOHN Timothy, LI Kevin A., GRISWOLD William G. *A Diary Study of Mobile Information Needs*,CHI'2008, Florence (Italie), 2008, [10 p.].

[STO 05] STOJANOVIC Nenad. *On the role of a user's knowledge gap in an information retrieval process*. In : K-CAP '05 Proceedings of the 3rd international conference on Knowledge capture, 2005, pp. 83-90.

[SME 99] SMEATON Alan F. *Using NLP or NLP Resources for Information Retrieval Tasks* In : Strzalkowski Tomek. Natural Language Information Retrieval, Kluwer Academic Publishers, 1999, p.99-111.

[SPA 99] SPARCK JONES Karen. *Whait is the role of the NLP in text retrieval ?* In : STRZALKOWSKI Tomek. Natural language information retrieval, 1999, pp.1-24.

[SPA 88] SPARCK JONES Karen. *A look back and a look forward*. In : SIGIR'88. Proceedings of the 11th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval. New York : ACM, 1988, pp.13-29.

[SPI 02] SPINK Amanda, OZMULTU H. Cenk. *Characteristics of question format web queries : an exploratory study*, INIST, vol.38 n.4, pp.453-471, 2002.

[SPI 01] SPINK Amanda, WOLFRAM Dietmar, JANSEN Bernard J., SARACEVIC Tefko. *Searching the web : The public and their queries*, Journal of the American Society for Information Science (JASIST), 2001, vol 52, n.3, pp. 226-234.

[STR 08] STROHMAIER Markus, PRETTENHOFER Peter, LUX Mathias. *Different Degrees of Explicitness in Intentional Artifacts : Studying User Goals in a Large Search Query Log*. In : In CSKGOI'08 International Workshop on Commonsense Knowledge and Goal Oriented Interfaces, in conjunction with IUI'08, 2008, [10 p.]

[STR 99] STRZALKOWSKI Tomek, LIN fang, WANG Jin, PEREZ-CARBALLO Jose. *Evaluating natural language processing techniques in information retrieval*. In : STRZALKOWSKI Tomek. Natural language information retrieval, 1999, pp.113-145.

[SUE 82] SUEUR Jean-Pierre. *Pour une grammaire du discours : élaboration d'une métjode : exemples d'application*, MOTS, vol.5, n.5, pp.143-185.

[SUI 13] SUIGNARD Philippe, KERROUA Sofiane. *Utilisation de contextes pour la correction automatique ou semi-automatique de réclamations clients*. In : Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013), 2013, Les Sables d'Olonne, France, p. 699-706.

[TAG 92] TAGUE-SUTCLIFFE Jean. *The pragmatics of information retrieval experimentation (revisited)*, Information Processing and Management, 1992, vol.28,n.4, pp.467-490.

[TAM 10] TAMINE-LECHANI Lynda, BOUGHANEM Mohand, DAOUD Mariam. *Evaluation of contextual information retrieval : overview of issues and research*, Knowledge and Information Systems, 2010, vol. 24, pp. 1-34.

[TAM 08] TAMINE-LECHANI Lynda, BOUGHANEM Mohand, ZEMIRLI Nesrine W. *Personalized document ranking : Exploiting evidence from multiple user interests for profiling and retrieval*, Journal of Digital Information Management, Digital Information Research Foundation (DIRF), 2008, vol. 6, n.5, pp.354-365.

[TAM 03] TAMINE-LECHANI Lynda, BOUGHANEM Mohand, CHRISMENT Claude. *Accès personnalisé à l'information : vers un modèle base sur les diagrammes d'influence*, revue I3 (Information, Interaction, Intelligence), 2003, vol.6, n.1, [22 p.].

[TAN 06] TAN Bin, SHEN Xuehua, ZHAI ChengXiang. *Mining Long-Term Search History to Improve Search Accuracy*, International Conference on Knowledge Discovery and Data Mining, 2006, pp.718-723.

[TAN 05] TANNIER Xavier, GIRARDOT Jean-Jacques, MATHIEU Mihaela. *Utilisation de la langue naturelle pour lŠinterrogation de documents structurés*. In : Actes CORIA 2005, [16 p].

[TIM 05] MITCHELL Timothy J.F., textscChen Sherry Y., MACREDIE Robert D. *Hypermedia Learning and Prior Knowledge : Domain Expertise vs. System Expertise*, Journal of Computer Assisted Learning, 2005, vol. 21, Issue(1), pp. 53-64.

[TAY 91] TAYLOR Robert. *Information use environments*.In : Brenda Dervin, Melvin J. Voigt (sous la dir.)Progress in communication sciences, 1991, vol.10, pp.217-255.

[THI 05] THIVANT Eric, BOUZIDI Laid. *Les pratiques d'Šaccès à l'Šinformation : le cas des concepteurs de produits de placements financiers*, Revue Electronique Suisse des Sciences de l'ŠInformation (RESSI), août 2005, pp.7-34.

[THI 01] THIVANT Eric. *Vers une modélisation des pratiques dŠaccès à lŠinformation*. Rapport de Recherche dans le cadre de la coopération franco-tunisienne en Sciences de l'Information et de la Communication, coopération avec l'Institut Supérieur de documentation (ISD), de l'Université de Tunis en 2001.

[TRE 03] TREIMAN R., CLIFTON C. Jr., MEYER A. S., WURM L. H. *Language comprehension and production. Comprehensive Handbook of Psychology*, vol. 4 : Experimental Psychology. New York : John Wiley and Sons, 2003, pp. 527-548.

[TRI 04b] TRICOT André. *La prise de conscience du besoin dŠinformation : une compétence documentaire fantôme ?*, site Doc pour Docs.

[TRI 03a] TRICOT André. *Apprentissage et recherche d'information avec des documents électroniques*. Mémoire pour l'habilitation à diriger des recherches, Université de Toulouse le Mirail, 128 p., 2003.

[TEE 07] TEEVAN Jaime, ADAR Eytan, JONES Rosie, POTTS Michael A.S. *Information re-retrieval : repeat queries in Yahoo's logs*. In : Proceeding SIGIR'07, 30th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 151-158 .

[VEI 85] VEILEX Florence. *Approche expérimentale des processus humains de compréhension en vue d'une indexation automatique des résumés scientifiques : application à un corpus de géologie*, thèse de doctorat, Grenoble 2, 1985, 156 p.

[VEN 09] VENTRESQUE Anthony, CAZALENS Sylvie, LAMARRE Philippe, VALDURIEZ Patrick. *Enrichissement sémantique de requêtes utilisant un ordre sur les concepts*. In : EGC ?2008 [10 p.]

[VOO 94] VOORHEES Ellen M. *Query expansion using lexical-semantic relations*. In : Actes 17th International ACM-SIGIR. Conference on research and development in information retrieval (SIGIRŠ94). New York : ACM Press, 1994, pp.61-69.

[WAE 04] WŒN Annika. *User Involvement in Automatic Filtering : An Experimental Study*, User Modeling and User-Adapted Interaction, 2004, vol. 14, Issue 2-3, pp. 201-207.

[WIL 96] WILSON Tom, WALSH Christina. *Information behaviour : an interdisciplinary perspective*. In : Proceedings of an International Conference on research in information needs, seeking and use in different contexts, Vakkari Pertti, Savolainen Reijo et Dervin Brenda, dir., Taylor Graham, pp. 39-49.

[WIL 81] WILSON Tom. *On user studies and information need*s, Journal of Documentation, 1981, vol. 37, Issue 1, pp. 3-15.

[WU 13] WU Wensheng, ZHONG Tingting. *Searching the Deep Web Using Proactive Phrase Queries*. In : Proceedings WWW'13 Companion, 22nd international conference on World Wide Web companion, 2013, pp.137-138.

[XU 13] XU Juan, ZHANG Qi, HUANG Xuanjing. *Understanding the Semantic Intent of Domain-Speci ?c Natural Language*. In : Proceedings International Joint Conference on Natural Language Processing, 2013, pp. 552-560.

[XU 96] XU Jinxi, CROFT Bruce W. *Query expansion using local and global document analysis*. International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.4-11, 1996.

[YAN 07] YANBE Yusuke, JATOWT Adam, NAKAMURA Satoshi, TANAKA Katsumi. *Can Social Bookmarking Enhance Search in the Web ?* . In : Proceedings of the JCDL'07, Conference on Digital Librairies, 2007, pp.107-166.

[YIN 10] YIN Xiaoxin, SHAH Sarthak. *Building taxonomy of web search intents for name entity queries*. In : Proceeding WWW'10 , 19th international conference on World wide web, 2010, pp.1001-1010.

[YVO 07] YVON François. *Une petite introduction au Traitement Automatique de la Langue*, notes introductives d'un cours sur le traitement des langues naturelles. ParisTech, 2007, [24 p.].

[ZHA 07] ZHANG Wei, LIU Shuang,YU Clement, SUN Chaojing, LIU Fang, MENG Weiyi. *Recognition and Classification of Noun Phrases in Queries for Effective Retrieval*. In : Proceeding CIKM'07, sixteenth ACM conference on Conference on Information and Knowledge Management, 2007, pp. 711-720.

ANNEXE I

GRAMMAIRE DES DEMANDES EN LN

D'après les travaux de [HUA 09], le point d'entrée de cette chaine d'analyse est le [REFERENT] que nous désignons z_a et la requête que nous désignons z_b, z étant la désignation du besoin informationnel. Si ce dernier s'exprime par un ou plusieurs mot-clé(s) strictement identique *if string1 == string2* alors nous la désignons comme correspondance totale *puts "match"* sinon nous *else, puts "no match*" effectuons les comparaisons désignées ci-dessous. Nous reprenons ces travaux de [HUA 09] que nous décrivons dans un langage formelle de notation.

Désignons un tiret bas (*underscore*) _ comme un espace de caractères. La ponctuation est représentée par un P compris dans les trois petits points (...) gauche de la requête. L'apostrophe, le tiret, le caractère numérique de datation sont exprimés de la façon suivante $P = ', ?, .$. La chaîne vide est représentée par λ. Désignons le $\sum$ comme l'alphabet des lettres, ds chiffres et de la ponctuation $\sum = \{[a-z], [0-9]\} \cup P$. c_i est un caractère de cet alphabet $c_i \in \sum$, w_i est un mot de cet alphabet $w_i \in^*$, et z_i est n'importe quelle chaîne composée de cet alphabet ou un espace de caractère $z_i \in (\sum \cup \{\_\})^*$, incluant la chaîne vide.

Ponctuations et espaces Nous traitons ici les ponctuations et espaces mais nous les considérons et donc les comptabilisons comme "correspondances totales".

$$z_{a1}\, v_1\, z_{a2} \overset{WP}{\rightarrow} z_{b1}\, v_2\, z_{b2} \tag{I.1}$$

$$v_1, v_2 \in P \cup \{\lambda, \_\} \tag{I.2}$$

$$\texttt{if any} \begin{cases} z_{a1}\, z_{a2} = z_{b1} z_{b2} \\ z_{a1}\, z_{a2} \overset{WP}{\rightarrow} z_{b1} z_{b2} \end{cases} \tag{I.3}$$

Ex : Auchan, swot => Auchan swot

Correction orthographique ou *Spelling Correction* (SC) Une correction orthographique est détectée en utilisant une fonction de distance de Levenshtein [LEI 96]. Cette fonction s'applique bien à ce cas présent car elle permet de faire correspondre le [REFERENT] et la requête même si ces derniers présentent des erreurs de frappes, inversions

de caractères et les caractères manquants. Le [REFERENT] et la requête sont classés comme correction orthographaphique si la distance Levenshtein est de 2 ou moins. Ex : accident automobile => accident automoible

Ré-ordonnancement de termes ou *Word Recorder* (WR) Les mots sont réorganisés mais restent inchangés :

$$z_a \overset{WR}{\rightarrow} z_b \tag{I.4}$$

$$if\ any \begin{cases} z_a = z_{a1\_}z_{a2} \quad , \quad z_b = z_{a2\_}z_{a1} \\ z_a = z_{a1\_}w\_z_{a2} \quad , \quad z_b = z_{b1\_}w\_z_{b2} \quad , z_{a1\_}z_{b2}, z_{a1\_}z_{a2} \overset{WR}{\rightarrow} z_{b1\_}z_{b2} \end{cases} \tag{I.5}$$

Ex : Japon Champagne => Champagne Japon Ici ; le [REFERENT] et la requête contiennent tous les deux les mêmes termes mais inversés. La première condition est le cas de base et la seconde condition est l'étape récursive.

Suppressions de termes : ou *Remove Word* (RW) Un ou plusieurs terme(s) est/sont supprimé(s) entre le REFERENT et la requête sur la base d'autres termes en commun.

$$z_a \overset{RW}{\rightarrow} z_b \tag{I.6}$$

$$if\ any \begin{cases} z_a = z_b \\ z_a \overset{WR}{\rightarrow} z_b \\ \_z_{a\_} = z_{a1\_}w_{x\_}z_{a2}, \_z_{c\_} = z_{a1\_}z_{a2}, z_c \overset{RW}{\rightarrow} z_b \end{cases} \tag{I.7}$$

Ex : boulangerie traditionnelle => boulangerie

Ajouts de termes ou *Add Words* (AW) Un ou plusieurs terme(s) est/sont ajouté(s) entre le REFERENT et la requête sur la base d'autres termes en commun.

$$z_a \overset{AW}{\rightarrow} z_b \texttt{ iff } z_b \overset{RW}{\rightarrow} z_a \tag{I.8}$$

Ex : bijouterie => horlogerie bijouterie

Racinisation ou stemming(Stem) Cette approche cherche à rassembler les différentes variantes d'un mot autour de son *stem* (*i.e.* une pseudo-racine). Cette procédure traite à la fois des cas relevant de la flexion et de la dérivation. Les techniques utilisées pour procéder à la racinisation reposent généralement sur une liste d'affixes[1] et sur un ensemble de règles de désuffixation construites *a priori*. Elles permettent de retrouver le *stem* d'un mot. L'avantage de ces outils (*stemmer*) réside dans leur simplicité : ces outils gèrent en même temps les morphologies dérivationnelle et flexionnelle. Le stemming est d'ailleurs la méthode la plus utilisée dans les moteurs de recherche [LOU 04]. Les règles du stemming sont décrites à partir de l'algorithme de Porter [POR 80] pour l'anglais ; elles consistent en sept phases successives et une cinquantaine de règles applicables. L'algorithme de Porter a été adapté en français par [PAT 02] avec l'algorithme de désuffixation Carry. Tout comme l'algorithme de Porter, Carry se décompose en phases successives. C'est également le suffixe le plus long qui détermine la règle à appliquer. Ainsi, des demandes portant sur des études de marché sur des «chiennes de race», le mot « chiennes » deviendra « chienn » par suppression du « s », du « e » final, puis « chien » par suppression de la double consonne finale.

$$w_{a1}\cdots\_w_{ai}\cdots w_{an} \overset{stem}{\rightarrow} w_{b1}\cdots\_w_{bi}\cdots w_{bn} \tag{I.9}$$

$$\texttt{if}\,\forall_i\Big(P(w_{ai}) = (Pw_{bi})\Big) \tag{I.10}$$

Ex : chiennes de race => chien de race

Acronymes ou *Form Acronym* (FA) Une transformation sous forme de sigle se produit lorsque la requête est un acronyme formé à partir des mots du [REFERENT].

$$c_1w_1\cdots\_c_iw_i\cdots c_nw_n\cdots \overset{\text{FA}}{\rightarrow} c_1\cdots c_i\cdots c_n \tag{I.11}$$

Ex : télévision => TV

Dans le même registre, une extension de l'acronyme ou *Expand Acronynm* (EA) se produit lorsque le [REFERENT] est un acronyme et que la requête contient les termes qui le composent.

[1] Un affixe est un morphème en théorie lié qui s'adjoint au radical ou au lexème d'un mot.

$$c_1 \cdots c_i \cdots c_n \overset{\text{EA}}{\rightarrow} c_1 w_1 \cdots c_i w_i \cdots \_ c_n w_n \tag{I.12}$$

Ex : LLD => location longue durée

Segments de chaînes ou *Substring* (Sub) Un segment de chaînes *Substring* est définie comme une instance où la requête est un préfixe ou un suffixe strict du [REFERENT]. Contrairement à la définition traditionnelle du segment de chaîne, cela ne comprend pas les cas où seuls les caractères à l'intérieur de la première requête sont extraits.

$$z_a z_b \overset{sub}{\rightarrow} z_a \mid z_b \tag{I.13}$$

Ex : étude sur l'ésotérisme => étude sur l'éso

Dans le même registre, *supertring* (Super) est défini comme une instance où la requête contient le [REFERENT] comme préfixe ou suffixe.

$$z_a \overset{Super}{\rightarrow} z_x z_a \mid z_a z_x \tag{I.14}$$

Ex : étude sur la para => étude sur la parapharmacie

Abréviations ou *Abbreviation* (Abbr) Une reformulation de l'abréviation est effectuée quand les termes du [REFERENT] et de la requête sont des préfixes de l'autre. Cela diffère des segments de chaîne *Substring* et *supertring* qui considèrent les suffixes et préfixes seulement. La reformulation de l'abréviation peut être détectée sur la totalité des requêtes.

$$w_{a1} \cdots \_ w_{ai} \cdots \_ w_{an} \overset{Abbr}{\rightarrow} w_{b1} \cdots \_ w_{bi} \cdots \_ w_{bn} \tag{I.15}$$

$$\texttt{if}\ \forall_i \left(w_{ai}\, w_c = w_{bi} \vee w_{ai} = w_{bi}\, w_c \right) \tag{I.16}$$

Ex : marché des déodorants parfumés => marché des déo parfumées

Substitutions de termes ou *word substitution* (WS) Le remplacement d'un mot se produit lorsque un ou plusieurs mots dans le [REFERENT] sont remplacés par les mots sémantiquement liés, déterminées à partir des thésaurus internes de la société (présentés à la partie 5.3. Deux termes sont liés si l'un a une relations sémantique (synonyme, hyponyme, hyperonyme, méronyme) avec l'autre. Cette règle est mise en œuvre en deux étapes : soit s'effectuer sur une partie seulement des termes soit sur la totalité des termes.

Prenons l'opérateur $\approx$ comme représentation sémantique entre deux termes, incluant la case lorsque les termes sont les mêmes.

$$z_a \overset{WS}{\rightarrow} z_b \tag{I.17}$$

$$z_a = w_{a1} \cdots w_{ai} \cdots w_{an} \tag{I.18}$$

$$z_b = w_{b1} \cdots w_{bi} \cdots w_{bn} \tag{I.19}$$

$$\texttt{if} \; \forall_i \left(w_{ai} \approx w_{bi} \right) \vee z_a \approx z_b \tag{I.20}$$

Ex : pain => baguette, hôtel => Kyriad ou encore fruits => banane

ANNEXE II

ÉTIQUETTES XELDA

Les étiquettes (*tags*) morpho-syntaxiques Xelda utilisées pendant la thèse sont présentées ci-dessous :

TAG	DESCRIPTION	EXEMPLES WORDS
+Abbr	abbreviation	ea.
+Accron	acronym	USA
+Adj	adjective	blue
+Adv	adverb	today
+Aux	auxiliary (verb)	will
+Bus	business name	Xerox
+Card	cardinal number	ten
+City	city name	London
+Comma	punctuation comma	,
+Comp	comparative	better
+Conj	conjuction	because
+Continent	continent name	Europe
+Country	country name	Scotland
+Day	day of the week	Monday
+Dec	decimal (number)	123
+Deg	(academic) Degree	PhD
+DlrAmt	amount of dollars	$
+Farm	last name	Clintons
+For	foreign language	c’est la vie
+Indef	indefinite (determiner)	a
+Inf	verb infinitive	be
+infto	infinitive particle	to
+Init	initial (proper name)	A.
+Interj	interjection	Oh
+Masc	masculine gender (proper names)	Peter
+Meas	mesurement unit	mm
+misc	miscellaneous place name	Thames
+Month	month name	January
+Nom	nominative case (pronoms)	I
+NomObl	nominative or oblique cas (pronouns)	you
+Non3sg	not third person singular (verbs)	play
+Noun	common noun	computer
+Num	spelled-out number	three
+Ord	ordinal (number)	me
+Paren	parentheses	(
+PastPerf	past participe	gone
+PastTense	past tense	went
+Percent	percentage	99%
+Pers	personal (pronoun)	us
+Pl	plural	houses
+Place	place name	Everest
+Poss	possessive (pronoun)	hers
+Prep	preposition	at
+Pres	present tense	play
+Prop	proper name	Xerox
+Pron	pronoun	nobody
+Punct	punctuation mark	;
+Quant	quantifier	some
+Quote	quotation mark	’
+Sent	sentence final punctutation	.